Data beyond the newsroom

Exploring the impact of data work in news media organizations

---

Universität Hamburg
Fakultät für Wirtschafts- und Sozialwissenschaften

Dissertation
zur Erlangung der Würde eines Doktors der
Wirtschafts- und Sozialwissenschaften (Dr. phil.)
gemäß der PromO vom 18. Januar 2017

Vorgelegt von Paul Solbach
Hamburg, 2024

Vorsitz: Prof. Dr. Judith Möller

Erstgutachterin: Prof. Dr. Wiebke Loosen

Zweitgutachterin: Prof. Dr. Juliane Lischka

Datum der Disputation: 14. Februar 2025

Eigenständigkeitserklärung

Hiermit erkläre ich, Paul Solbach, dass ich keine kommerzielle Promotionsberatung in Anspruch genommen habe. Die Arbeit wurde nicht schon einmal in einem früheren Promotionsverfahren angenommen oder als ungenügend beurteilt.

Eidesstattliche Versicherung

Ich, Paul Solbach, versichere an Eides statt, dass ich die Dissertation mit dem Titel „Data beyond the newsroom. Exploring the impact of data work in news media organizations" selbst und bei einer Zusammenarbeit mit anderen Wissenschaftlerinnen oder Wissenschaftlern gemäß den beigefügten Darlegungen nach § 6 Abs. 3 der Promotionsordnung der Fakultät für Wirtschafts- und Sozialwissenschaften vom 18. Januar 2017 verfasst habe. Andere als die angegebenen Hilfsmittel habe ich nicht benutzt.

Hamburg, Juni 2024

"When you cannot measure it, when you cannot express it in numbers,
your knowledge is of a meagre and unsatisfactory kind."
— *Lord Kelvin*[1]

"What gets measured, gets managed."
— *Peter Drucker*[2]

[1] Often misquoted as "If you cannot measure it, then it is not science." (Kelvin, 1889)
[2] While this adage is commonly attributed to the prominent management educator Peter F. Drucker, it might not have been said by Drucker verbatim (Zak, 2013). In any case, the full (alleged) quote reads: "What gets measured gets managed—even when it's pointless to measure and manage it, and even if it harms the purpose of the organization to do so." (Barnett, 2015)

Table of Contents

List of Tables & Figures

1. Introduction

Working with data, and the notion of workable data, has seen explosive growth at digital news publishers in recent years. While editorial analytics and data journalism have generated considerable research interest, the quality and impact of data work *outside* of the newsroom remains comparatively obscure. With this thesis, I trace the phenomenon back to its origins, describe organizational changes and challenges around it, and show how it affects the field of journalism overall. To this end, I propose a two-fold approach. First, a closer look at how data, metrics, and data affordances are implemented and evaluated across media organizations. Second, by examining emerging data roles and data practices inside these organizations, I intend to investigate shifting or even blurred boundaries in a traditionally antagonistic system between the editorial center and the ancillary publishing machine around it. Ultimately, I want to reach a deeper understanding of the notion of data work inside the field.

Building on a theoretical framework of complementary concepts—from organizational isomorphism at the inter-organizational level, through boundary objects on the case level, and a professionalized data gaze on the level of individual data workers—the empirical part of this thesis consists of six case studies conducted at German news media organizations of various sizes and legacies, from a regional newspaper brand to a digital-native online publication. Overall, the aim is to contribute fruitful empirical input to the growing body of data work research and provide a useful analytical framework for further study of data work in journalism and beyond.

1.1 Why study data work?

As Silicon Valley investor Marc Andreessen famously put it, "software is eating the world" (Andreessen, 2011). A deluge of data is generated and fed back into the expansive universe of software in a circular fashion: software generates data, more software is built or optimized as a consequence of data. All material aspects of our social interactions, business conduct, governance, consumption, and knowledge are replicated into digital form (Bucher, 2018; Deuze, 2012; Fuller & Goffey, 2012; Marres, 2017; Ruppert, Law & Savage, 2013). Rightfully, some have declared an entire "data revolution" underway (Kitchin, 2022). While these assertions have now become commonplace and datafication as pushed by the likes of *Google* and *Meta* is now the subject of public discourse, there are less spectacular but every bit as fascinating datafication processes that fundamentally change our environment. Inertial industries and public sector institutions are still in their datafication infancy, with data and data work a subset of an ongoing general digitalization.

Why does the study of data work in news businesses matter? The implications of digitalization are not yet fully understood, and they are constantly re-evaluated by media organizations, platforms, and academic researchers alike. Digital platforms tackle this uncertainty and rapid change as "perpetual experiment engines" (Crawford, 2014) that approach society in constant beta-testing mode (Marres, 2017; Neff & Stark, 2002). Measuring performance data, usage data, interaction data, competitive data and even data about data (system monitoring data, metadata) seems to have developed into a prerequisite for survival in a new era of data Darwinism where datafication is at times portrayed as a "political-economic regime" (Sadowski, 2019) that continues to capture markets until all markets have been subdued to its power.

While prominent news brands were early adopters of an emerging World Wide Web in the 1990s, methodical experimentation efforts in the sense of rapid product iteration and innovation only picked up in the 2010s as economic pressure reached a breaking point (Flew, 2012; Westlund & Lewis, 2014; Lewis & Usher, 2013)—with systematic operational data work, as we will see, only becoming a notable development in recent years. Against this backdrop, studying datafication and data work appears especially fruitful in the context of news. Are there "data regimes" (in a critical data studies sense) evolving inside news organizations and if so, do these organizations adopt counter-practices as a consequence? What exactly is the role of data scientists in news businesses? Can data work transcend the paradoxa (Loosen, Pörksen & Scholl, 2008, p. 23) between the economic interests of the publishing house and the qualitative standards of the editorial staff and serve as a valuable resource for both parties?[3]

Data plays a four-fold role in digital news, as a source of reporting, as a topic of reporting, as a means of communication and as an enabler of operations management. By contrast, journalism studies on datafication are largely concerned with editorial perspectives, with some notable exceptions (Bechmann, Bilgrav-Nielsen & Korsgaard, 2016; Evens & Van Damme, 2016). As editorial analytics have played an integral part in the now-diminishing business model of display advertising, research on the topic is plentiful (Tandoc, 2014; Tandoc, 2019; Cherubini & Nielsen, 2016). In recent years, data journalism[4] and the role of data in news production has gained attention (Cushion, Lewis & Callaghan, 2017).

[3] The challenge of balancing social responsibility with the economic interests of media companies extends into digital news organizations as well. Herman and Chomsky's propaganda model (2010) provides a critical examination of this issue.

[4] For a discussion on the various -isms in journalism, see (Loosen et al., 2022).

News personalization has been discussed as a problem of diversity (Haim, Graefe & Brosius, 2018; Bodó, Helberger, Eskens & Möller, 2019), serendipity (Fletcher & Nielsen, 2018) and in terms of algorithmic accountability and transparency (Van Drunen & Helberger, 2019). All these approaches replicate the inherent logic of the field, with editorial concerns as the pivotal point of all news production. Given the ubiquity of data and automatization, it seems reasonable to direct more attention to organizational changes outside the editorial realm. Data and data infrastructure not only constitutes the basis for personalization, dynamic paywalls, or user research. Arguably, contrary to the implications of the term "raw data", a seemingly pure and unadulterated form of data never exists (Gitelmann, 2013). As data represents decisions, opinions and abstractions made more or less intentionally, extra care needs to be taken in considering the conditions under which data are generated and proliferated in the context of news businesses.

Given the intensive public and scholarly interest in artificial intelligence or machine learning (ML), which by definition requires large amounts of well-formed training data (Miceli & Posada, 2022), I will also explore the practical relevance and requirements of this technology in daily news business or ask where it might provide utility in the future. Often linked to this particular technology, data science is not limited to the modelling and operationalization of data in the form of advanced ML. With data science roles emerging at news publishers, I aim to better understand their relationship with ML—though not as a main objective.

Finally, in the interest of disclosure and to better explain my personal motivation, I want to share some details about my professional background.

After pursuing a career in journalism as a reporter, I started at the R&D unit of Deutsche Presseagentur in 2013—during a particularly dynamic phase of innovation-driven journalism. With social media competing for audience attention, digital-native news companies such as Vice and Buzzfeed News outperforming legacy media[5], and continuing pressure to adapt to new technologies (from 360-degree imaging to chatbots, podcasts, video and augmented or virtual reality), the only constant was change. Our unit engaged in cooperation with other research institutions like Fraunhofer IIS to work on big data streams of text and audio and acted as part of the international committee on news schema standardization[6]. Discussions surrounding data were ubiquitous. After my time at Deutsche Presseagentur, I founded a media technology startup which helped news organizations better understand their audiences through data collection and dashboards. Overall, the notion of data and its various aspects, including infrastructure, metadata, proprietary data, structured data, big data and data fluency, emerged as a prevalent theme of innovation within the field. While I share the belief in data professionalism as a requirement for most digital businesses to succeed, I have long had the desire to take a more empirical look at the role of data and data work in organizations to substantiate (or shake) such a belief.

[5] A good indicator of how the news landscape changes drastically, both these companies have fallen into economic decline since then: Vice Media filed for bankruptcy, Buzzfeed News shut down entirely (Darcy, 2023; Hirsch & Mullin, 2023)

[6] IPTC, International Press Telecommunications Council

1.2 Attending to data work

As made evident by the multi-faceted nature of data, its political, technological, organizational, and social dimensions, a scientific approach to the phenomenon of data work needs to cover a lot of ground. This is acknowledged by the nascent field of *critical data studies*, understanding data as an "amalgam of ideas, methods, technologies and stakeholders" (Kitchin, 2022, p. 23). Such an amalgam has been theorized as a *data assemblage*, a complex socio-technical system composed of many apparatuses and elements that are thoroughly intertwined, whose central concern is the production, management, analysis, and translation of data for a particular purpose (Lauriault, 2014). Sociologically speaking, data is needed to make claims, shape and control processes, and the asymmetries in data power have far-reaching consequences. Consequently, data power, or how agenda setting and decision-making are enforced by powerful and instrumented data and metrics that are increasingly collected and governed by corporations (Taylor & Broeders, 2015), should be another key concern of critical data studies.

Examining changes at news organizations by looking at data work, I need to first gather a suitable conceptual framework. With critical data studies as an overarching "research theme" (Selwyn, 2022, p. 594), I suggest a combination of three theoretical layers. At the outmost layer, I intend to examine my cases through the lenses of collective rationality and *organizational isomorphism* as developed by DiMaggio and Powell (1983). As will be explored in greater detail, isomorphism refers to rationalization, bureaucratization and other forms of organizational change as the "result of processes that make organizations more similar without necessarily making them more efficient" (DiMaggio & Powell, 1983, p. 147). While these assertions were made many years ago, they remain especially robust in the context of more recent work on algorithmic and data-

driven technology (Caplan & Boyd, 2018). With data work here (assumingly) representing the economic intentions of marketplace participants, these lenses seem ideally suited to my study. While my empirical scope does not allow for any claims about the overall effectiveness of data work, I can tend to the imaginations and reasonings related to data efforts at news organizations. It remains to be shown if, and to what extent, the field of digital news businesses constitutes a "context in which individual efforts to deal rationally with uncertainty and constraint often lead, in the aggregate, to homogeneity in structure, culture, and output" (DiMaggio & Powell, 1983, p. 147). Can I observe that data work and data processes unfold in similar ways across organizations? Are these simply organizations responding to the responses of other organizations (Schelling, 1978)? Where do they vary and for what reasons? If I encounter homogenous data structures and practices, then are these the result of coercive forces, a mimesis in the face of great uncertainty or simply a growing professionalization within the field?

One level down from this overarching research theme, I then approach the specifics of data work inside news organizations. At this level, we need to touch on all stages of the data lifecycle: from data generation, handling, processing, storage, sharing and analysis to interpretation and even the question of data removal. A key theoretical concept in critical data studies, Kitchin and Lauriault (2014) propose to consider data assemblages as made up of two elements: the *technological stack* and the *contextual stack*. While the description of specific technological stacks appears to be relatively straightforward, with data infrastructure made up of certain storage, transformation, and interaction affordances whose characteristics are available to us both publicly (e.g. in the form of software documentation or corporate blogs) as well as through expert interviews, the contextual stack is more diffuse.

In their pursuit of authority over particular definitional domains, professions engage in boundary maintenance to varying degrees (Abbott, 2014; Carlson & Lewis 2015; Gieryn, 1983; Lamont & Molnar, 2002; Star, 1989). To illuminate the contextual stack, it might be productive to view data workers as interloper or boundary workers who advocate and reinforce fields of knowledge. Such an approach has proved fruitful in examining the shifts between traditional and emerging fields in professional journalism (Lewis & Usher, 2016; Belair-Gagnon & Holton, 2018; Eldridge, 2019). Following along the conceptual path of professional boundaries, we might then explore the properties and utilities of data artifacts such as dashboards and reports, as they are envisioned by data workers and data managers. How do perspectives on these technical affordances differ? Could these "boundary negotiating artifacts" (Lee, 2007) help to enforce certain rules and managerial strategies, renegotiate, or further entrench boundaries inside news organizations?

Finally, we need to consider the individual agency of data workers (Wu, Tandoc & Salmon, 2019). In the context of this study, I propose to adapt the notion of a professionalized and self-affirming *data gaze* (Beer, 2019). Rather pessimistically, this notion extends the concept of the *medical gaze* (Foucault, 2013), wherein patient bodies fall subject to manipulation by the professional authority of medical practitioners. In this analogy, data workers acquire the status of powerful data librarians, without whom the exegesis of data might be deemed impossible. As soon as organizations adopt the narrative of a data revolution, the argument goes, data power "lies firmly in the hands of those who are able to interpret or tell stories with the data" (Beer, 2016).

Coincidentally, the notion of the data gaze overlaps with the gatekeeping function often attributed to and claimed by journalists (Bro & Wallberg, 2014; Shoemaker, Vos & Reese, 2008). It appears fruitful to observe if and to what extent my material either confirms or conflicts with Beer's theses about the circumstances of the "codified clinic" (Beer, 2018, p. 87) or the self-perceptions of data workers.

In sum, understanding the interplay of data work and managerial goals would allow a better grasp on the transformation towards a datafied, metrics-driven and automated practice. By breaking down ways of thinking, responsibilities and metrics along the technological and contextual stacks, I aim to construct a clearer sense of how technological innovation affects the entire field. To this end, I propose to undertake exploratory case studies at a broad spectrum of digital news organizations large and small. Qualitative interviews should provide a solid understanding of emerging data roles, infrastructure, and metrics, as well as the negotiating power of data work inside highly flexible professional boundaries.

1.3 Thesis outline

The thesis starts from the recognition that new forms of data work are being done at news businesses which not only introduce new technologies and digital artifacts but potentially reshape the professional boundaries inside those organizations. This recognition prompts us to first consider what the social sciences have established about data and data work, the dualities inherent in news production processes, and only then, embarking on our empirical research. As mentioned above, my research questions are addressed empirically through exploratory case studies.

The thesis is organized as follows. In the introductory chapter, I outline my research project and make the case for why examining data work in news businesses is important and may offer a fresh angle on digital transformation in the industry. I also gather a theoretical reference frame that consists of three layers: from collective rationality and organizational isomorphism at the inter-organizational level, through data artifacts as boundary objects on the case level, and a professionalized data gaze on the individual level.

Chapter 2 introduces the theoretical reference points laid out above in greater detail. Chapter 3 provides a general overview of the elements of data and data work, definitional groundwork, and the state of research on data work from a multitude of fields and a closer look at the role of data, algorithmic systems, and automation in the newsroom. With the theoretical framework not entirely specific to data or data work, the aim is to supplement said theoretical framework with a clear understanding of data terminology and concepts as well as prior research around data in the field of journalism before I then specify guiding assumptions and research questions in the next chapter.

In Chapter 4, various guiding assumptions are constructed on the basis of the theoretical framework, which ultimately lead to a set of research questions. On the basis of these research questions, I develop a course of inquiry in the next chapter.

Chapter 5 illustrates the research design, triangulating qualitative data from semi-structured expert interviews that were carried out as part of exploratory case studies. After going into greater detail about the case selection criteria and methods used to conduct case interviews, I then give an overview of the sample resulting from the empirical phase of the study and explain the steps taken during analysis.

Chapter 6 contains the empirical results gathered through the coding and subsequent analysis of my qualitative data. For each case, I first provide a relational diagram of the roles and units involved with data and data work at the time of inquiry. Then, more detailed analysis is given across the topics of organizational (re-)structuring, data-informed decision-making, specific metrics employed inside the organization, perspectives on editorial requirements and general editorial input, as well as descriptions of technical data infrastructure. Lastly, I discuss my findings across all theoretical layers.

In Chapter 7, I formulate conclusions about what my study of data work in news businesses can bring to the field of journalism research and beyond. I summarize the main contributions of this dissertation and reflect on the challenges and implications of data work and sketch a few directions for future research. As we are dealing with technical terminology, industry jargon and abbreviations throughout the dissertation, the appendix contains a glossary of terms that might otherwise be unknown or unclear to readers.

2. Theoretical frame

2.1 Purpose of this frame

Drawing from my personal experience and knowledge working in the field and looking at previous data-related inquiries around editorial or web analytics (e.g. Tandoc, 2019; Petre, 2015), data-driven journalism (e.g. Howard, 2014; Gray, Chambers & Bounegru, 2012) and datafication/automation in journalism (e.g. Diakopoulos, 2019; Belair-Gagnon & Holton, 2018; Baack, 2015), I consider the phenomenon of data work in journalism to be sufficiently uncovered. In that sense, starting with a well-defined and relatively narrow substantive area of inquiry, theory here provides the guiding framework to both study design and questioning.

Following the presupposition above, we want to understand how data work beyond the newsroom has changed in the field, assert the factors leading to such developments, the specific qualities of data work and data affordances, and also understand how boundaries across departments are negotiated in terms of data work (beyond-ness implies some form of boundary). The goal then becomes to investigate how the self-perceptions and professional norms of the specialists might contribute to such a solidification or permeability of boundaries. Based on these general interests and avenues of inquiry, I selected a theoretical framework which starts broad at the top and narrows downwards, reflecting the transition from the meso to micro levels of theory.[7]

[7] Arguably, the theoretical layers presented here do not neatly fall into a single category. CDS spans multiple levels in its focus on individual instances of communication (micro-level) and power structures (meso-level) but it often uses these instances to comment on broader societal or institutional patterns (macro-level). The data gaze in turn also addresses broader social impacts of a data thinking (macro-level). I focus on the meso-level aspects of CDS and the micro-level aspects of the data gaze respectively.

At the broadest level, I capture the technical and political complexities of data work through the lens of critical data studies and then transition down to the organizational viewpoint with the help of collective rationality and organizational isomorphism, reflecting on how data work might stem from and contribute to patterns of similarity among organizations. A step closer to the ground, I scrutinize the tangible data affordances and artifacts, where data workers negotiate professional boundaries around boundary objects and at the same time shape the way data work gets carried out in their organizations. Finally, at the individual level, I examine the personal perspective of data workers through the notion of a professionalized *data gaze*. Such a layered approach, moving from the universal to the individual level, should provide a comprehensive lens towards data work from its broader implications to its most intricate details.

In this chapter, I lay out the framework in more detail and show how it points towards the method of qualitative expert interviews, providing the basis for "thick descriptions" (Geertz, 1973) of data and data work in the field.

2.2 From organizational to individual layers

2.2.1 Critical data studies

As I will discuss in Chapter 3, the notion of recorded data has been around for centuries. More recently, with widespread internet access and use, massive amounts of data are generated and acted upon with ever increasing levels of automation and scale—what is historically known as big data.[8] Rather than treat these large-scale data as merely empirical and objective phenomena, critical data studies (CDS) advocates that these data should be viewed as constituted within wider "data assemblages" and "data regimes"—the properties and entanglements of which I want to deconstruct in the field of digital news organizations.

Originating from the field of geography (Dalton & Thatcher, 2014; Graham, 2014; Kitchin, 2014), CDS applies critical social theory to data, exploring how they are never simply neutral, objective, independent, raw representations of the world, but are situated in, contingent on and relational to other things and "do active work in the world" (Kitchin & Lauriault, 2014). CDS to this day remains a relatively informal field, a "loose knit group of frameworks, proposals, questions and manifestos" (Illiadis & Russo, 2016, p. 1) or an "amalgam of ideas, methods, technologies and stakeholders" (Kitchin, 2022, p. 23). Kitchin and Lauriault (2014) theorize data assemblages, complex socio-technical systems composed of many "apparatuses and elements" (2014, p. 8) that are thoroughly intertwined, whose central concern is the production, management, analysis and translation of data for a particular purpose (2014).

[8] With the criteria of "bigness" a constant subject of debate, other scholars have suggested to focus instead on big data as redefining power dynamics involved in the processes of data production and knowledge discovery (Balazka & Rodighiero, 2020). I propose to discard the term altogether, supported by arguments laid out in Chapter 3.

In simpler terms, data is needed to make claims, shape and control processes, and the asymmetries in data power have far-reaching consequences. Consequently, critical data studies should also focus on data power, meaning how agenda-setting and decision-making are driven by influential data and metrics (Taylor & Broeders, 2015). To illustrate a complex data assemblage, Kitchin uses the example of population census:

> "[A census] is underpinned by a realist system of thought, it has a diverse set of accompanying forms of supporting documentation, its questions are negotiated by many stakeholders, its costs are a source of contention, its administering and reporting is shaped by legal frameworks and regulations, it is delivered through a diverse set of practices, undertaken by many workers, using a range of materials and infrastructures and its data feed into all kinds of uses and secondary markets." (Kitchin 2022, p. 25)

In turn, a census could then be regarded as part of a greater data ecosystem characterized by the interdependency between organisms and resources who are constantly "seeking equilibrium through motion rather than stasis" yet vulnerable to "exogenous forces which may disrupt or destroy the ecosystem" (van Schalkwyk, Willmers & McNaughton, 2016, p. 69). While clearly inspired by the thinking of Deleuze, Guattari and Latour, the original authors refrain from explicitly basing their notion of the assemblage on prior work. Overall, the various assemblage readings share a relational view of social reality in which human action results from shifting interdependencies between material, narrative and social elements. Here, an assemblage has the same general properties as other readings of the term, in that it consists of associations between humans and non-humans, both internal and external to the system, whose constellation is constantly changing. This notion of a data assemblage

is similar to Foucault's (Foucault, 1980a; Foucault, 1980b) concept of the *dispositif*, which refers to a "thoroughly heterogeneous ensemble of discourses, institutions, architectural forms, regulatory decisions, laws, administrative measures, scientific propositions, philosophical, moral and philanthropic propositions" (Foucault, 1980b, p. 194) that reinforce and sustain the exercise of power in society. In Foucaults thinking, the dispositif of a data infrastructure generates what he calls "power/knowledge"—knowledge that serves a strategic function. In that sense, data infrastructures are never neutral, essential, or objective, their data never raw, but always "cooked" (Gitelman, 2013) according to a particular recipe by "chefs embedded in institutions that pursue particular aspirations and goals and who operate within a larger framework" (Kitchin & Lauriault, 2014, p. 9).

As evidence of such powerful data assemblages, CDS points to the works of statistician and science philosopher Ian Hacking (1992, 1999, 2002, 2007), who in turn drew inspiration from Foucault's thinking on knowledge production. Hacking discusses many instances of "cooking" data (for instance around census recordings) and shows how statistical patterns can develop explanatory functions in and of themselves, making the work appear non-deterministic. Assuming interrelated processes within an assemblage of data which produce and legitimize the data and associated devices/elements, Hacking envisions these processes to shape the way data operates in the world—influencing future iterations of itself and the assemblage as a whole.[9] In these processes, he sees a "dynamic nominalism" (Hacking, 1992, p. 78) at work, where there is an interaction between data and what it represents that leads to mutual change.

[9] It is important to note how Hacking never expressly talked about "assemblages" of "data", let alone consider the implications of ("big") data affordances and how power dynamics might shift towards the providers and operators of such affordances.

There are other variations of this idea of reciprocity between the measurement and its target from which a socio-technological observer effect could be construed, but for the sake of brevity, I stick closely with the hypotheses gathered by CDS.[10] The main process Hacking dubs the "looping effect", a circular series of data classification, organization, ontology and actuation steps:

> *a) classification,* usually within a category, "a most general principle" of categorization; *b) objects of focus* (e.g., people, spaces, fashions, diseases, etc.), in case of *humans*, ascribed characteristics of groupings of items or entities eventually become part of their self-identity; while for non-human entities, people develop notions and interactions based on their classification *c) institutions*, who "firm up the classifications", or manage data infrastructures; *d) knowledge*, that deliberates and defines the characteristics which in turn constitute classifications; *e) experts* within administrations or institutions who hold said knowledge, tasked with the classification. (Hacking, 2007, pp. 288–289)

Having first formulated his theses in the 1980s, before the arrival of the internet, with population-level data collections in mind, Hacking then goes on to describe how these loops of steps result in the "making up of people" (Hacking, 2007, p. 285)—the classification loop works to "reshape society in the image of a data ontology" (Kitchin & Lauriault, 2014, p. 11). Nonetheless, Hacking's thinking remains largely adaptable to large-scale data regimes within private organizations.

[10] For instance, "Campbell's Law", after social scientist Donald T. Campbell, posits: "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor." (Campbell, 1979, p. 89)

For the present study, tracing my material for a circular system of data, objects, institutions, knowledge and experts seems fruitful, albeit without following the original concept of classification.[11] In a later addition to these socio-political considerations on data assemblages, Kitchin describes digital data systems as made up of two parts: the technical stack, "instrumental means by which data are generated, processed, stored, shared, analyzed and experienced" and the contextual stack, "a number of discursive and material components related to philosophy and knowledge, finance and politics, law and governance, practices, stakeholders and actors, geography and markets" (Kitchin, 2022, p. 23–25). While the former, technical stack appears to be an inherent part of data assemblages or regimes, the latter one seems infinite in scope. Here, I will demarcate my study to concern itself with the contextual stack of organizations. To approximate the data assemblage in my field of inquiry, I will infer a (non-exhaustive) mapping of apparatuses (systems of thought, forms of knowledge, practices, institutions, places) and its respective elements from my material.

As a sort of manifesto, the original advocates of CDS posit a handful of "provocations" (Dalton & Thatcher, 2014). Three of these I will embrace for the present study: 1) situate data regimes in time and space 2) expose data as whose interests they serve 3) illustrate the ways in which data are never raw. It is apparent how these "provocations" are rather abstract and could be seen as a cross-cutting concern of this study and thought of as enveloping more specific theoretical considerations.

[11] (Statistical) "classification" makes sense in the case of population census data, but appears too narrow in terms of large-scale data regimes in organizations. While classification lies at the root of many Machine Learning use cases (see also Kotsiantis, Zaharakis & Pintelas, 2007), I expect to find various other data problems in the cases examined (regression, clustering or even non-statistical uses).

2.2.2 Organizational isomorphism

Reflecting on Weber's analogy of rationalization as an "iron cage", sociologists Powell and DiMaggio in 1983 contended that organizations of a particular field, having achieved full bureaucratization, appear to grow increasingly similar without necessarily becoming more efficient (DiMaggio & Powell 1983). Structured organizational fields provide a context in which "individual efforts to deal rationally with constraint and uncertainty in aggregate lead to homogeneity in structure, culture and output." (DiMaggio & Powell, 1983, p. 147) Journalism qualifies as a field of organizations, as its organizations in aggregate constitute a recognized area of institutional life; and displays all four properties of structuration given by the authors: a) an increase in the extent of interaction among organizations in the field; b) the emergence of sharply defined interorganizational structures of domination and patterns of coalition; c) an increase in the information load with which organizations in a field must contend (Boczkowski, 2005; Anderson, 2013; Pavlik & Bridges, 2013) and d) the development of mutual awareness among participants in a set or organizations that they are involved in a common enterprise (shared performance metrics like impressions, users or circulation; Picard, 2011, pp. 59–71).

The scholars then posit three mechanisms of institutional isomorphic change and provide a set of predictors as to the isomorphic properties of any given organizational field. In the following, I will discuss all three mechanisms and a subset of these predictors I expect to apply to my field of inquiry. In examining data work and data regimes across multiple digital news organizations, I am particularly interested in the reasons for these organizations in the same field to coincide or drift apart with their specific approaches. Thus, the assumptions of organizational isomorphism theory are particularly relevant.

As part of new institutionalism, the theory of isomorphism can be included in a number of significant advancements in organizational sociology that have led to a resurgence of the field. Since the 1960s, empirical studies on organizational structures and processes had gathered systematic information on samples of organizations—either of the same type or of diverse types within the same area (Scott, 2004). This mode of operation continued through subsequent decades, but it was not until the 1990s that studies were conducted based on a representative sample of organizations in a single society (Kalleberg et al., 1996). These new types of studies, particularly those involving multiple types of organizations and societal contexts, affirmed a dualist nature of organizations—shaped "in part by material-resource forces, and in part by social and cultural systems" (Scott, 2004, p. 8). With the advent of New Institutionalism, sociology adopted a more differentiated view on decision making in institutions: from intentionally driven, rational choices on the basis of uncertainty, ambiguity, risk preference and conflict to a more nuanced process, with decision making as a vision driven by a "logic of appropriateness" (see March & Olsen, 2011), as constructed through an array of organizational rules and practices, not by a logic of consequence.[12] More succinctly, according to DiMaggio & Powell, the formal structures of an organization did not necessarily reflect rational or optimal ends, but were instead are "a matter of myth and ceremony" (Alvesson & Spicer, 2019, p. 2), creating the illusion of rationality and legitimacy.

[12] Especially interesting in this context are case studies on information technology adoption and implementation, starting with the work of Paul Attewell (Attewell, 1992). Attewell challenges the predominant emphasis on processes of influence and information flow for technology dissemination, and focuses on the relevance of know-how and organizational learning as potential barriers to adoption of innovations. Firms delay in-house adoption of complex technology until they obtain sufficient technical know-how to implement and operate it successfully.

As to the reasons behind these myths and ceremonies, the scholars identify three mechanisms that drive isomorphic change within a field: *coercion*, *mimeticism* and *normativity*.

On a political level, *coercive isomorphism* results from "formal and informal pressures exerted on organizations by other organizations upon which they are dependent and by cultural expectations in the society within which organizations function" (DiMaggio & Powell, 1983, p. 150). Examples of dependent organizations include regulatory and fiscal requirements enforced by the state onto legal entities. Secondly, organizational structures within a field may also coincide as a result of imitation.[13] Such mimetic processes are especially encouraged "when organizational technologies are poorly understood, when goals are ambiguous, or when the environment creates symbolic uncertainty" (DiMaggio & Powell, 1983, p. 151). The authors also refer to this mimesis as a process of modeling, where the modeled organization "merely serves as a convenient source of practices that the borrowing organization may use". Such organizational models can diffuse unintentionally, through personnel mobility or explicitly through consulting organizations or industry associations (DiMaggio & Powell 1983, p. 151).[14]

[13] As economist Armen Alchian put it: "While there certainly are those who consciously innovate, there are those who, in their imperfect attempts to imitate others, *unconsciously* innovate by *unwittingly acquiring* some unexpected or unsought unique attributes which under the prevailing circumstances prove partly responsible for the success. Others, in turn, will attempt to *copy the uniqueness*, and the innovation-imitation process continues." (Alchian, 1950, pp. 218–219)

[14] Mimetic isomorphism echoes the *cargo cult* phenomenon, as in the modeling of EU innovation policies after their successful US counterparts: "The key 'ritual' structures are increased R&D expenditures; an emphasis upon the commercialization of science through university-based spin-outs and licensing routes in high-technology producing sectors; the promotion of entrepreneurship and new business entry; and a supposed US entrepreneurial culture based on the subsidization of risk taking in venture capital investment and of the development of the SME sector more generally." (Hughes, 2010, p. 101)

Lastly, organizational structures in a field converge under normative pressures, meaning their convergence follows internal or external rules and regulations, but is not directly market-related:

> "Similarity can make it easier for organizations to transact with other organizations, to attract career-minded staff, to be acknowledged as legitimate and reputable, and to fit into administrative categories that define eligibility for public and private grants and contracts." (DiMaggio & Powell, 1983, p. 153)

While these operational gains are not directly quantifiable, normative convergence helps professionals in the field create a "recognized hierarchy of status, of center and periphery, that becomes a matrix for information flows and personnel movement across organizations" (DiMaggio & Powell 1983, p. 152). In this sense, while data and data work transport imaginaries of measurability and effectiveness, their adoption may as well be regarded as a way of signaling to the workforce—performative data work as status competition.

How could I operationalize the question of isomorphism? Among other aspects, assessing and discussing data assemblages (Chapter 2.2.1), the technical specifics and contextual stacks around data will provide a solid foundation for inter-organizational comparison. Additionally, I have selected a subset of predictors of isomorphic change I expect to positively apply to my material—discarding such predictors that consider inter-organizational dependence or resource constraints which do not match the purposes of this study[15]:

[15] For instance, while Hypothesis A5 addresses an issue also found within the field of journalism (and by extension data workers in journalism), it seems trivial in the context of this study: "The greater the reliance on academic credentials in choosing managerial and staff

Organization-level predictors

- The more uncertain the relationship between means and ends, the greater the extent to which an organization will model itself after organizations it perceives to be successful. (A3, mimetic)
- The more ambiguous the goals of an organization, the greater the extent to which the organization will model itself after organizations that it perceives to be successful (A4, mimetic)
- The greater the participation of organizational managers in professional associations, the more likely the organization will be, or will become, like other organizations in its field. (A6, normative)

Field-level predictors

- The fewer the number of visible alternative organizational models in a field, the faster the rate of isomorphism in that field. (B3)
- The greater the extent to which technologies are uncertain or goals are ambiguous within a field, the greater the rate of isomorphic change. (B4)
- The greater the amount of professionalization in a field, the greater the amount of institutional isomorphic change. (B5)
- The greater the extent of structuration of a field, the greater the degree of isomorphism (B6) [16]

personnel, the greater the extent to which an organization will become like other organizations in its field." See also Hartmann, 2018; Hanitzsch, 2019.

[16] DiMaggio & Powell, 1983, pp. 154—155

2.2.3 Boundary objects

Originally introduced by sociologist Susan Leigh Star in a 1989 article on distributed artificial intelligence[17], the notion of boundary objects centers around the idea of artifacts shared between disparate groups to facilitate collaboration, "both plastic enough to adapt to local needs and constraints of the several parties employing them, yet robust enough to maintain a common identity across sites. They are weakly structured in common use, and become strongly structured in individual-site use". (Star, 1989, p. 46) Star developed the concept after examining how scientists of heterogenous backgrounds collaborate successfully, despite not having "good models of each other's work" with "different audiences to satisfy" and employing different "units of analysis, methods of aggregating data, and different abstractions of data" (Star, 1989, p. 46).

In her original proposal, intended as an impulse rather than a complete theory, Leigh asserted a taxonomy of boundary objects she found in her studies. First, "repositories" are such boundary objects, that are "indexed in a standardized fashion" (Star, p. 47). These can be physical knowledge bases, file registers and libraries, as well as digital objects like present-day spreadsheets, wikis, or databases—with a nomenclature of "indexes", "unique identifiers" and "object-relationality", databases appear in close proximity to what Star had in mind.[18] Secondly, "platonic" objects are abstractions such as maps, which are distant enough from particular domains to work for all of them (with maps not

[17] Star, S. L. (1989). The structure of ill-structured solutions: Boundary objects and heterogeneous distributed problem solving. (pp. 37–54).

[18] To be precise, the concept of "index" means something else than in e.g. the semiotics of Charles Sanders Peirce, where an index shows a physical relationship with its reference and points towards its meaning. As a concept of object-relational databases, "indexes" enumerate the entries of a "table", the main constituent of a whole database.

representing the actual physical world while capable of holding information from multi-disciplinary inputs). Platonic boundary objects provide the advantage of adaptability at the cost of local contingencies. Thirdly, "coincident" boundary objects are objects from different domains that show agreement in their boundary formulation, while zooming in on various overlapping focus areas driven by various goals. Finally, more a specification of the "repository" object, Star thinks of standardized "forms" as deleting local contingencies to allow for comparability (e.g. patient symptom forms in clinical studies).

Building on these ideas, the device of boundary objects has been adopted by computer science (e.g. Subrahmanian, Monarch et al., 2003), information science (Huvila, Anderson et al., 2016), education theory (e.g. Akkerman & Bakker, 2011), and management studies (e.g. Benn & Martin, 2010; Spee & Jarzabkowski, 2009; Koskinen & Mäkinen, 2009)—showing how boundary objects aid in enforcing managerial strategy or maintaining coherence across intersecting social worlds of management. In journalism studies, boundary objects have been shown as transcending the editorial "firewall" (Perreault, Kananovich et al., 2022; Coddington, 2015), allowing for different meanings to be drawn from news articles across social domains (Scott, Bunce et al., 2019), enabling interlopers or boundary workers like web analytics providers (Belair-Gagnon & Holton, 2018), enabling experiences through "networked witnessing" (Ananny, 2015), or aiding in the negotiation of entirely new boundary demarcations, for instance, between editors and software developers (Lewis & Usher, 2016). Across disciplines, thinking in boundary objects has been a particularly fruitful approach in the context of case studies (Bergermann & Hanke, 2017).

Over time, the concept has grown and evolved. Notably, Charlotte Lee proposed that we consider periods of unstandardized and destabilized organization where objects are transient and changing, which she coins as "boundary negotiating artifacts" (Lee, 2007). Prompted by this widespread application and extension of her work, Star herself stated how "much of the use of the concept has concentrated on the aspect of interpretive flexibility and has often mistaken or conflated this flexibility with the process of tacking back-and-forth between the ill-structured and well-structured aspects of the arrangements."

Honoring these words of caution, I will take extra care to not linger on data itself as a data boundary object too much—a somewhat trivial assertion, with data inherently stateful, standardized, meant to be passed on, enriched across domains etc., but more on the specifics of how data objects contain and represent power in the sense of CDS, how they might reveal asymmetries across the participating domains, how and if data objects are fostering a "milieu of constant experimentation" (Belair-Gagnon & Holton, 2018, p. 14), and/or a constant renegotiation of boundaries. Overall, Star's original article leaves some potential untapped in a) not discussing the staggering conceptual parallels to semiotics and b) presenting a fuzzy taxonomy to begin with. Still, based on the numerous findings and discussions of boundary fluidity in the field of journalism, I expect to generate a useful discussion around my sample data as well.

2.2.4 The data gaze

Descending further down the theoretical framework, from the outermost layer of social theory provided by CDS, isomorphism, and boundary objects on the organizational level, we arrive at the individual level of data workers as actors with a certain agency. Based on the observation that the ways in which journalism is currently undergoing change do not simply happen, but are also explicitly, avowedly, and more or less purposefully driven by actors, I will discuss my material in light of the data gaze—a theoretical lens developed by sociologist David Beer on the basis of Foucault's concept of the professional medical gaze—which the philosopher first introduced with the "Birth of the Clinic" in 1973.

Foucault develops his notion of the medical gaze as a corollary of the modern clinical practice at the turn of the 18th century when medicine began to focus on the observation and examination of the physical body, a period in which "the whole dark underside of disease came to light, at the same time illuminating and eliminating itself like night, in the deep, visible, solid, enclosed, but accessible space of the human body" (Foucault, 2003, p. 195). What was fundamentally invisible inside the human body, was "suddenly offered to the brightness of the gaze" (Foucault, 2003, p. 195). During this era, understanding of disease shifted from a focus on symptoms as reported by the patient to the physical signs as anatomically observed by the physician. Foucault then casts the observant clinician's work as a form of rule-based data processing and describes the clinician's gaze as "directed upon a succession and upon an area of pathological events; it had to be both synchronic and diachronic, but in any case it was placed under temporal obedience; it analyzed a series." (Foucault, 2003, pp. 162–163).

In collecting patient data, making observations about the human body, running diagnoses with the impetus of increased methodological rigor, medical professionals aimed to develop an objective or neutral way of understanding the body. But according to Foucault, their medical gaze was instead governed by the sociopolitical and historical context in which it operated. As a result, enabled by the setting of the modern clinic, medical practitioners at the same time increasingly reinforced existing power dynamics, as well as marginalized certain groups of people. While considered by Foucault a product of the enlightened bourgeoisie (Foucault, 2003, p. 74), characterized by a focus on scientific knowledge and the use of objective methods to understand the body, the medical gaze has often led to pathologize and stigmatize groups such as women, people of color and people with disabilities (Plsek & Greenhalgh, 2001). By extension, Foucault illustrates how such a new form of knowledge also produces a new type of measured language, "for the dream of an arithmetical structure of medical language must be substituted, therefore, by the search for a certain internal measurement consisting of fidelity and fixity, of primary and absolute openness to things and rigour in the considered use of semantic values." (Foucault, 2003, pp. 114–115). Describing facts then emerges as the "supreme art in medicine" and "everything pales before it", Foucault concludes (Foucault, 2003, p. 115). On a mission towards a manifestation of truth, the medical gaze aims for an "exhaustive description" (Foucault, 2003, pp. 113–114) of its subjects, not of the totality of the human body—but of the details needed to prescribe (Foucault, 2003, p. 100, 196). Foreshadowing the promises of automated activity on the large-scale measurements in digital spaces, a mythical "speaking eye" then becomes the imagination of such translation from objective measurement into signs and signifiers, into language, by a neutral entity hovering over the clinic:

> "It would scan the entire hospital field, taking in and gathering together each of the singular events that occurred within it; and as it saw, [...] it would be turned into speech that states and teaches the truth, which events, in their repetitions and convergence, would outline under its gaze, would, by this same gaze and in the same order, be reserved, in the form of teaching, to those who do not know and have not yet seen. This speaking eye would be the servant of things and the master of truth." (Foucault, 2003, p. 115)

Lastly, the philosopher extrapolates from the historical formation of clinical medicine, seeing it as merely one of the more visible witnesses to the changes in fundamental structures of experience and even going so far as to reconcile its positivist character with phenomenology in that it already contained "the original powers of the perceived and its correlation with language in the original forms of experience, the organization of objectivity on the basis of sign values, the secretly linguistic structure of the datum, the constitutive character of corporal spatiality, the importance of finitude in the relation of man with truth, and in the foundation of this relation, all this was involved in the genesis of positivism" (Foucault, 2003, p. 199).

Many parallels to data work and the assumptions made by CDS about data and data professionals start to unfold in Foucault's thinking—with medicine an example of a highly skilled and seemingly data-informed, science-led profession that might exert social power under the premise of total objectivity.[19]

[19] Medicine has historically legitimized racist and discriminatory policies and practices. For instance, in the 19th and early 20th centuries, many doctors and scientists used their medical knowledge and expertise to argue for the superiority of certain races and to justify policies that were designed to oppress and discriminate against others. This included the use of pseudoscientific concepts such as "eugenics" or "phrenology" to support policies of forced sterilization and segregation. See also Weikart, 2016; Galton, 1904.

Consequently, Beer extends the concept of the medical gaze to the profession of data workers and data service providers, also employing Jacques Derrida's image of archons, powerful record-keepers, who oversee the storage and retrieval of data and metadata as soon as it accumulates, and who ultimately "have the real sway" (Beer, 2019, p. 12).

The data gaze gets conceptualized in four parts, starting with the general assertion of how it appears inextricably linked to emergent intermediaries or service providers (specifically data analytics providers), who facilitate the circulation of data and instill the gaze as external forces to an existing economy and in doing so increase their influence. Beer assets how these providers spread their influence by "both the analytics that they provide and also with the way in which they theorise, represent and project power onto data." (Beer, 2019, p. 38) Secondly, temporality shapes the gaze in the sense of a "need to accelerate so as to keep up with the accelerating world" (Beer, 2019, p. 42), with speed not only a countermeasure against inefficiency and waste but the conception of real-time data promising the "possibility of reacting quickly, gaining an edge, winning the competition and even anticipating future events" (Beer, 2019, p. 48). Third, in recourse to Foucault, the data gaze cannot operate outside data infrastructure to host its data professionals, the analytical space of what Beer calls the "codified clinic" (Beer, 2019, p. 81). These codified clinics are painted by its beneficiaries as a complex ecosystem, appear to always be under review and development, only fully graspable through deep expertise and insider knowledge. Lastly, the diagnostic eye of data workers, like data analysts and engineers, are said to embody the data imaginary as they are "expected to translate, to render digestible, to find value" in data (Beer, 2019, p. 122) while also supervising the codified clinic and keeping it running.

Although reflection on the concept appears very fruitful for the present study, a few caveats and disagreements need to be addressed: tasked with data collection, diagnostics, analysis, and agency, medical practitioners may be akin to other data workers in this regard, yet their decisions cross over into the physical with far-reaching consequences—a difference Beer himself acknowledges multiple times (Beer, 2019, pp. 9, 132). The amount of power exerted by computational data workers has yet to be proven in the context of my material. Second, as indicated by the rather extreme basic premise, Beer appears quite biased on the subject matter of data, as made evident by numerous examples: the author's data imaginary gets constructed on the basis of a newness, importance and scale that is said to be claimed by the imaginary itself. Third, and most relevant to the present study, the basic assertion of how service providers and, by proxy, the data imaginary are something external to their customers, needs to be challenged.

Overall, Beer paints a dystopian view of data and data work, painting sane companies as being infiltrated by data service providers, which he sees as instrumental in creating a "black-box society" (Beer, 2019, p. 32). In some places, the author's apparent disdain for the data and analytics industry even leads to mischaracterizations of technological concepts and nomenclature.[20] Having raised my objections here to an otherwise stimulating concept, I will proceed to incorporate it into my questioning.

[20] As an example, Beer states how the term "nested data", which simply means data structures containing one another, has a suggestive undertone: "[The term] catches the eye in this passage and is suggestive of this creation of a safe and secure space in which to hold the data until they are used." (Beer, 2019, p. 78) Seeing as the origins of the term "nested" can be traced back as far as the late nineteenth century (describing the practice of putting one container within another for storage or transport; Annual Iowa State Report, Vol. 12, 1889) it certainly was not invented as modern-day "marketing speak".

3. State of research

3.1 Overview

In the structure of a more theory-oriented qualitative study like the present one, both theory and literature discussion might often be located in separate sections toward the beginning of the write-up (Creswell & Creswell, 2018, p. 68) or help in situating the need for research in the context of a case study design (Creswell & Poth, 2018, p. 185). While literature discussion will take place in the findings section, I follow said approach and supplement my theoretical framework with a clearer understanding of concepts, terminology and prior research around data in this chapter, before I then specify guiding assumptions and research questions in the next. Overall, literature and research review might also help identify gaps either in the current state of research or my own line of thinking.

Starting with clarifications around the basic concepts of data, information, and knowledge, I then show how the concept of data work appears closely linked to the concept of knowledge work and where it finds application in other fields like medical or education studies—although I am unable to cover all shapes and forms of data work discussions across other industries and cultural spheres. Finally, I will take a closer look at media and journalism studies on the general topic of data and datafication to substantiate why I consider the phenomenon of data work to be generally uncovered in the field but still lacking a more recent perspective looking beyond the newsroom.

3.2 Aspects of data

Originating from the Latin word *datum*, which means "something given", the term data emerged in the 17th century (Kitchin, 2022, p. 5).[21] The Latin root indicates how data can be regarded as provided or given to someone for a specific purpose—as opposed to something materially pre-existent. In today's Cambridge dictionary, data is defined as "information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer".[22] A closely related and equally fuzzy concept, "information" as a term first appeared in the 14th century (Gellert, 2022, p. 158), linked to both Latin and French origins.[23] In the mind of economist Fritz Machlup, information refers to "telling something" or to "something that is being told", it is about the transmission and reception of messages (Machlup, 1983, p. 660). Such a grasp of information makes it intrinsically linked to the notion of messages and messengers (see also Capurro, 2009). It becomes apparent how both concepts share the qualities of directedness, of activity towards something. But how exactly is data different from information? Distinction between the two concepts remains a matter of passionate debate, with information scholars offering rather esoteric constructions: "Data can refer both to sensory stimuli and to a set of signs that represent empirical stimuli. Information refers to both the meaning of stimuli, and to a set of signs which represent empirical knowledge." (Gellert, 2022, p. 159) According to other sources, a common differentiation would be data as the unprocessed representation of facts or

[21] Scholarly dispute around the use of the syntactical plural data as a collective noun can be traced back over centuries (Keen & Blake, 1927) and is still ongoing (Kitchin, 2022). I will assume the syntactical plural for the present study.

[22] Cambridge Advanced Learner's Dictionary, 2013.

[23] The term is linked to both Latin and French, combining "inform", meaning to give a form to the mind, and the activity suffix "-ation". As such, its literal meaning could be interpreted as "the training, or molding, of the mind" (Logan, 2012, p. 70).

observations, while information represents data organized into meaning, relevance and usefulness (Buckland, 1991, pp. 351–352). While regularly operationalized this way, the notion of data as factual has been largely outdated. Among others, media scholars have pointed out how data should always be looked at as "cooked" according to a particular recipe (Gitelman, 2013; D'Ignazio & Klein, 2020).[24] But with data ill-positioned as the basis for information, what does that say about the knowledge society? As this brief etymology indicates, the term data is enigmatic and today takes on many forms, intentions or meanings depending on the context. To better understand what data means and arrive at a clear conception for our field of inquiry, let us first go back in time and trace the historical contexts of data and working with data.

In ancient civilizations, early forms of data collection and epistemic record-keeping emerged as means to document important information. Examples include cuneiform tablets in Mesopotamia, hieroglyphs in ancient Egypt and quipus in the Inca civilization (Coulmas, 1989). As precursors to the cuneiform script, humans used elaborate methods of tallying and documenting commodities with clay tokens. As linguistics scholars posit, this evolution of writing from crude shapes to ever more elaborate scripture, illustrates nothing less than "the development of information processing to deal with *larger amounts of data* in ever *greater abstraction*" (Schmandt-Besserat, 2014; Emphasis added). From the outset, data appears inextricably linked to the phenomenon of humans writing, recording and materializing data in ever greater detail and quantity while at the same time, inventing abstractions of the data itself—a data paradox. These early writing systems served a host of

[24] It makes sense to consider this notion as "post-disciplinary"—a term Silvio Waisbord uses to point out how "disciplinary boundaries are fluid" to media and communication studies, a field not "interested in defining and patrolling epistemological boundaries". (Waisbord, 2019)

practical purposes, such as recording transactions, organizing administrative tasks and preserving cultural and historical knowledge—but data required another considerable leap to become the object of scientific measurements we think of today, which is the birth of statistics. Antecedents of statistics can be traced back to ancient societies such as the Babylonians or later China's Han Dynasty (Durand, 1960) maintaining census data for tax collections, population surveys or military purposes. While these practices certainly laid the foundation for statistical thinking, it was not until the 17th century that a field of statistics with its own modes of analysis truly emerged (Stigler, 1986).[25] During the Scientific Revolution, scientists had elevated systematic observation and experimentation around quantitative data to the condition of scientific reasoning. In Isaac Newton's mathematical way of working (on mechanics and optics), measurements and rules "provided the quantitative data and formulas upon which mathematical demonstrations in physical sciences depend" (Strong, 1951, p. 91). In the aftermath of these paradigmatic changes, first statistical methods like least squares (Stigler, 1986, p. 4) or complete theories of probability (Shiryaev, 2016, pp. 14–15) could flourish. With the technological advancements of the Industrial Age[26], manufacturers began to gather and analyze extensive data on production, transportation and macroeconomic trends. Such a data-driven approach gave companies the "constant flow of information essential to the efficient operation of these new large business domains" (Chandler, 1978, p. 109) and control through statistics quickly became both a science and an art for the managerial class of the time. This "visible hand" of management replaced market mechanisms as the core

[25] Stigler considers earlier examples of statistics such as the 12th century *Trial of the Pyx* as "isolated instances of human ingenuity" that failed to be "developed and integrated into a formal scientific discipline". (Stigler, 1986, p. 3)

[26] Arnold Toynbee's labeling and conception of the British Industrial Revolution remains controversial, as it encapsulates the approach of economic history and gets used very liberally by historians for different purposes (see also Wilson, 2014; Hoppit, 1987).

developmental and structuring impetus of modern business (Chandler, 1978, p. 455). In this way, statistics and quantitative data could be considered a side effect of the industrious organization around machines. Data and information evolved into a primary constituent of the 20th century, as exemplified by the notion of the Information Age, a discourse originating in the 1960s, when futurists, policymakers, journalists, social scientists, and humanists started writing about the "coming of a new era based on computers and communications technology" (Kline, 2015, p. 5). After successfully infiltrating businesses and organizations, the information processing model became the quintessential way of describing how individuals made decisions and the brain converted its inputs into its outputs, its stimulus into response—a thinking encapsulated by the contemporary field of Cybernetics (see also Kline, 2015). Although social scientists heavily criticized this dominant discourse around information in the 1980s[27], phrases like information economy, information society, or information age have been woven into everyday speech ever since. Undeniably, with the development of "computerized record-keeping systems" (Lucey, 2004, p. 32) such as databases and spreadsheet software, data became more accessible and available. Yet these terms contained a certain belief that interconnected information creates a new, desirable economic and social order.

[27] "Every historical period has its godword. There was an Age of Faith, an Age of Reason, an Age of Discovery. Our time has been nominated to be the Age of Information. [...] Unlike 'faith' or 'reason' or 'discovery', information is touched with a comfortably secure, noncommittal connotation. There is neither drama nor high purpose to it. It is bland to the core and, for that very reason, nicely invulnerable. Information smacks of safe neutrality; it is simple, helping heaping up of unassailable facts. In that innocent guise, it is the perfect starting point for a technocratic political agenda that wants as little exposure for its objectives as possible. After all, what can anyone say against information?" Roszak, 1994, p. 19.

The rise of the internet and digital communication channels further expanded the scope and scale of data collection, as increasing amounts of networked devices generate increasingly more data—nothing less than the "datafication of the world" (Filipović, 2015, p. 6)[28]. With my field of inquiry pertaining to data collection at digital media organizations, I will discuss the current invocation of data and big data through the lens of large-scale *datafication* in the next chapter. To conclude this chapter, we have traced how data carried a quality of directedness at various stages through history, with bureaucratic power structures seeking to organize societies and organizations, decision making under uncertainty by managerial data practitioners all inherent in the concept of data and information.

[28] Quote translated from German. All non-English quotes in this thesis are translated into English by the author.

3.3 Large-scale datafication

First introduced in 2013 as "the ability to render into data many aspects of the world that have never been quantified before" (Cukier & Mayer-Schönberger, 2013, p. 29), datafication is commonly associated with the collection, databasing, quantification and analysis of information as a source for knowledge production, service optimization and economic value-generation. While often attributed to the rise of digital technologies, cloud infrastructure and new types of networked devices like smart sensors, datafication could also be considered as the continuation of ancient practices—as was demonstrated in the previous chapter. Such a reading of the term roots it historically in practices of population management and the emergence of modern forms of the bureaucratic organizing of states and companies (Porter, 1996), assisted by old media technologies such as writing and printing to register or administer people and society (Cukier & Mayer-Schönberger, 2013). Tax, health, educational, and social security records in the welfare state were historically kept in written documents (Dencik & Kaun, 2020); and so was productivity and performance measurement of workers (Pollard, 1965) or customer information in retail (Turow, 2017). In short, datafication precedes digitalization. Yet following the digitalization of the Global North through internet infrastructure and access since the 1990s, such operations of collecting and processing information have become both unprecedented in scope and scale as well as increasingly automated (Andrejevic, 2014). Today, datafication marks how digital systems fuel, intensify, and automate historical practices of databasing, analyzing, and using information as a key resource for value creation—instilling these practices into everyday life. It is this notion of datafication as a mass phenomenon that I want to discuss in more detail.

Datafication research has seen contributions from a wide array of disciplines, most prominently in media and communication, education, and sociology.[29] Within these disciplines, research subjects of datafication studies are mostly twofold, either concerning user perspectives or infrastructure with a predominantly meso-level scope of analysis (Flensburg & Lomborg, 2021). Infrastructural studies are skewed towards studies of organizations and sectors in society, asking questions pertaining to how infrastructures work in various organizations and institutions (Andrew, 2019) or how datafication processes transform industries and business models (Nieborg & Poell, 2018). Aspects of datafication have also been the subject of various literature reviews in recent years. For instance, Kennedy et al. (2021) map "original empirical research into public understanding and perceptions of, attitudes towards and feelings about data practices and related phenomena" (Kennedy et al., 2021, p. 3), an empirical angle that has flourished lately according to the authors. Coming from another angle, Ruckenstein & Schüll (2017) review datafication in the context of health and identify two attitudes to data: "The so-called big-data fundamentalists promote the view that large data sets, properly mined for correlations and patterns, will render up previously elusive insights, predictions, and answers to long-standing challenges of individual and collective life, replacing the need for theory and science" (Ruckenstein & Schüll, 2017, p. 262) On the other hand, they see social science and humanities scholars promoting "a more skeptical stance, emphasizing the cultural, political, economic, and rhetorical dimensions of the data paradigm shift, typically by focusing on particular cases of 'datafication', or the conversion of qualitative aspects of life into quantified data" (Ruckenstein & Schüll, 2017, p. 262).

[29] Based on a literature review of a dataset comprised of 463 publications on datafication (See also Flensburg & Lomborg, 2021).

As examples of critical datafication perspectives, scholars such as Beer (2018) and Hepp (2019) predominantly attribute datafication to expansive forces embodied by certain actors from the technology spectrum. I believe these works at times inadvertently perpetuate their own narrative of growth where more nuance and objectivity might be advisable. As Hepp asserts, "we can observe that datafication has a dual character [...] the term not only captures a trend in the sense of changes that have already occurred but it also manages to encapsulate expectations of its own stability and growth." (Hepp, 2019, p. 49) From the perspective of digital infrastructure, datafication refers "first and foremost to the existence of appropriate data centers that enable the centralized collection and processing of data in the cloud" (Hepp, 2019, p. 48). However important the discourses on platform governance, anti-trust or the commodification of users might be, datafication should not be conflated with these topics. For example, according to Cisco (2021), global increase in internet traffic, "the exabytes of data being transferred to and from cloud servers" (Hepp, 2019, p. 62) does not directly correlate with an increase in the collection of sensitive user data but rather with the proliferation of (ultra-high-definition) streaming content from video-on-demand platforms, while the data flow of quantitative user data is miniscule in comparison (or at least does not grow proportionally).[30] In my mind, datafication as the "rendering into data" of new things and the rather trivial assertion of how (redundant) data throughput keeps growing should be two separate discussions.

[30] A report by global networking hardware corporation Cisco shows a threefold growth in internet traffic between 2016 and 2021 with a threefold increase of video content, video making up 84% of consumer IP traffic in 2021. A solid argument for growing datafication might be a massive growth in the number of networked devices per individual, allowing for more detailed user profiles (Cisco Systems, Global 2021 Forecast Highlights). Boyd & Crawford make similar arguments in that the "quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem" (Boyd & Crawford, 2011, p. 1).

In terms of its theoretical underpinnings, sociological work on datafication has been "greatly shaped" (Couldry, 2020, p. 1135) by actor-network theory (ANT).[31] As this study centers around data and datafication processes, we will touch briefly on ANT fundamentals, its critique, or outright dismissal by datafication scholars. A groundbreaking novelty of (what came to be labeled) ANT lies in the idea of actors emerging from the constitution of a heterogeneous network, where nodes are not only social but also material things, non-social things and artifacts (the "missing masses"; Latour, 1992), or knowledge imbued with agency. Reflective of this notion, the term "actants" takes the place of "actors" to indicate that agency is not only attributed to human beings. Actor-networks emerge as the result of transformation processes in which the activities and characteristics of all involved actants are incorporated and changed. As such, actants are not predefined in their networking function but are brought forth through the process of networking. As equivalent agency is awarded to persons and things, ANT scholars have denied the intentionality of either, reducing purposeful action and intentionality to "properties of institutions and apparatuses" (Latour, 1999, p. 192). ANT sees itself as a constructivist approach beyond social constructivism (Latour & Woolgar 1986, p. 281), where *macro-phenomena* (stable and delineated actor-networks) are explained as emergent from contextualized *micro-processes* (Peuker, 2010, p. 326). Some datafication scholars have cast doubt on the adequacy of this framing (Couldry, 2020) and how its focus on smaller organizational units would "obstruct us from grasping emergent forms of platform power and the new scale of social processing which they are helping to generate" (Couldry, 2020, p. 1146).

[31] As developed mainly by Michel Callon, Bruno Latour and John Law (Peuker, 2010, p. 325). Latour himself rarely used the term himself, calling ANT "more a method to deploy the actor's own world building activities than an alternative social theory" (Latour, 1999, p. 19).

Instead, in the sense of grand theory building, only a critical perspective towards processes of social formation and social order would allow the necessary scope to deal with the transformative impacts of large internet platforms that increasingly impact our lives (Couldry, 2020).[32] Addressing the shortcomings perceived in ANT's flat ontology[33], the central question of the datafied society should then become: "How is the overall order of social life being reconfigured to promote particular corporate and governmental interests on the basis of new and radical forms of reduction—the reduction of human life to configurations from which profit through data can be maximally extracted?" (Couldry 2020, p. 114).

For the present study, I acknowledge the "profoundly ideological role" (Van Dijck, 2014, p. 5) of individual beliefs that are embedded in data power. But as my interest lies in the organizational level of data work, I will steer around the wider debate and presuppose datafication as a remarkably modernist narrative and mode of operation, in a way still pursuant of the industrious organization around machines. Within the ambition of grand theory building, datafication gets painted as a continuously expansive societal phenomenon from the outset. But on an organizational level, datafication might turn out to be a limited or even reversible process.

[32] As an alternative approach, three theoretical supplements are proposed to approach datafication: the concept of figurations developed by Norbert Elias; Luc Boltanski and Judith Butler's contrasting accounts of definitional and categorical power; and the social theory of capitalism developed by Karl Marx and Moishe Postone; Couldry, 2020, p. 1145; Postone, 1993.

[33] "Rather than treating one type of object such as quantum particles as the really real upon which all else is grounded and to which all else ultimately reduces, flat ontology advocates a pluralism of types of objects at all levels of scale that are irreducible to one another. In other words, objects of different types and at different levels of scale are what Aristotle referred to as genuine primary substances." Bryant, 2011, p. 280.

3.4 Working with data

With a clearer understanding of data, its etymology, societal impact, and current discourse around the phenomenon of datafication, we continue with the concept of data work. While I will produce my own definition of data work in the context of our field of inquiry (following the empirical results), we need to gain an overview of previous conceptions first. Historically, data work appears closely linked to the idea of the knowledge worker, first conceptualized in the late 1950s as individuals whose primary contribution to an organization lies in the application of intellectual capabilities to solve complex problems, make decisions, and innovate—unlike traditional workers who relied primarily on physical labor.[34] Arguably, this paradigm shift was also anticipated by Marx, according to whom knowledge becomes an important productive force as technological productivity grows: "The development of fixed capital indicates to what degree general social knowledge has become a direct force of production, and hence, to what degree the conditions of the process of social life itself have come under the control of the general intellect and been transformed in accordance with it. To what degree the powers of social production have been produced, not only in the form of knowledge, but also as immediate organs of social practice of the real-life process." (Marx, 1858, p. 706) Present day data workers are clearly part of a more specialized practice that involves elaborate data systems and tooling. This inherent quality of working with digitized data and computing was already imbued in data work as the term first appeared in the context of information studies and systems theory in the 1980s, with studies conceptualizing hierarchies of information work as containing knowledge producers (in the narrow meaning of specialists

[34] In "Landmarks of Tomorrow" (originally published in 1959), Peter Drucker also characterized knowledge workers ("people doing knowledge work") by their ability to adapt according to the priciples and concepts of automation (Drucker, 1996, p. 67).

conducting research, solving complex problems and developing innovations), and data workers, jobs involving the processing and the application of knowledge produced by the producers, using systems and tools developed by the producers (McLoughlin, Rose & Clark, 1985). Again, the distinction between the modes of knowledge production and a subordinate data working class bears a striking resemblance to the principles of Marxian thinking.[35]

We find various examples of research on data work as a primary objective in medical studies (Fiske, Prainsack & Buyx, 2019; Pedersen & Bossen, 2021; Møller, Bossen et al.; 2020; Pedersen, 2022), education studies (Lu & Dillahunt, 2021; Foster & McLeod, 2018) as well as human-computer interaction (Miceli & Posada, 2022; Rothschild, Meng et al., 2022; Hockenhull & Cohn, 2021; Feinberg, Sutherland et al., 2020) and to a lesser extent in information management or computer science (Sambasivan, Kapania et al., 2021). As for the reasons why studies on data work as a practice are predominantly found in medical and education studies, I assume it might be related to a) the systemic importance of the underlying professional domains which makes it imperative to run constant second-order diagnostics and b) a track record of catastrophic measurements and classifications in these fields.[36] In the field of human-computer interaction, data workers are understood as a community of practice, where "members of the periphery receive less attention as compared to full practitioners, e.g. data scientists" (Rothschild, Meng et al., 2022, p. 307). Across the theoretical literature used for the present study, data work either gets discussed off-handedly or not at all.

[35] More commonly, knowledge workers as a whole (regardless of the specific function of producing or analyzing information) are portrayed as the post-modern working class in the Marxian sense (Fuchs & Mosco, 2015; Fuchs, 2014). Other voices consider knowledge work as representative of a quaternary sector of the economy (Kenessey, 1987).

[36] Radical eugenics (Galton, 1904) in medicine; highly consequential statistical advantages in early education (Merton, 1968; Gladwell, 2008).

Despite being a central figure in CDS, Kitchin does not use the term in his comprehensive monograph on data (Kitchin, 2021). Beer, on the other hand, refers to data workers multiple times to broadly signify all staff involved with data, not limited to analysts but extending to e.g. data engineers (Beer, 2019, p. 119). In recent non–academic writing, data workers are identified as the "exploited labor" (Williams, Miceli & Gebru, 2022) behind artificial intelligence.

A cursory search on Google Scholar turns up 34,300 exact matches for the term data work. In comparison, other common data bigrams produce vastly more results.[37] Overall, data work could be said to not hold a clear definition and as such it does not carry nearly the same discursive weight as other data-related terms like business intelligence or data science. Yet these terms were developed in the field, suggesting a superior precision and truthfulness of their making, as opposed to the much more neutral and inclusive data work.[38] I expect organizations to use data towards operational goals, to mold and manipulate data. In a sense data are actively and intentionally "worked into" processes and cultures, as I will demonstrate. Such a directedness and intentionality of activities around data finds expression in the "working" aspect of the term as well. A conception of data work as carried out by data workers, not merely as a fixed but a transitory role taken on by management and general employees alike, seems ideally suited to encompass all functions, practices and intentions related to data in organizations.

[37] Over a million results for "data science" (August 2023). These numbers are only an indication: a 2-gram (a sequence of two words) makes it difficult to generate empirically valid results here, problems of disambiguation and anaphora resolution are evident in queries like these (Mitkov, 2014).

[38] "Business intelligence" is often attributed to Hans Peter Luhn, a researcher at IBM, who discussed the concept of "Business Intelligence Systems" in a 1958 article as a precursor to general data processing and analysis in businesses. "Data science", on the other hand, has its foundations in academia, but was appropriated by the technology industry in the late 2000s (Davenport & Patil, 2012).

3.5 Data and the news

As I have claimed at the outset of this study, data has progressively become a focal point of discussion within journalism and journalism studies in recent years. To further substantiate this claim, I will discuss three broader topics of research within digital journalism studies, outlining key findings in order of their appearance. First, I will look at research on editorial analytics. Second, I discuss data journalism, where data serves as a means of uncovering, supporting, and communicating news stories. Finally, I will look at how computational methods and related professional classes shape the journalistic news production cycle. As we progress through each category, I aim to cast some light on the multifaceted relationship between journalism practitioners and data work and its implications for this study. Notably, with very few exceptions, the cited research centers around data-related phenomena inside the confines of the newsroom.

First, we need to position the various data discourses against the backdrop of a general state of journalism in terms of digitalization and technological change. Research on datafication in newsrooms finds innovation through data primarily occurring and becoming apparent in the following four areas: the visualization of large amounts of data, through, for example, interactive graphics or dashboards; altered editorial structures, often carried out by data teams and data experts; new narrative formats, including digital storytelling and multimedia formats; as well as general influences on topic selection due to the real-time analysis of news consumption behavior (Schätz & Pühringer, 2022). Again, I expect to add a broader cultural perspective to these findings by opening the questioning to roles beyond the newsroom. We also find a strong critical discourse on the perceptions of innovativeness (Subramanian & Nilakanta, 1996) inherent in the field of journalism.

Seminal research around innovation, entrepreneurship and startups in the field finds innovation commonly understood as a "novelty or change that is typically associated with an improvement, advancement or progress in journalism" (Buschow & Wellbrock, 2020, p. 8) with the latter criterion often implicitly assumed. Other researchers ascertain how an increasingly datafied environment simply demands "innovative processes and techniques for filtering and presenting relevant information" (Schätz & Pühringer, 2022, p. 19, translated by the author). What are these processes and techniques in particular? Given the increasing challenge of misinformation on social network sites (SNS), fact-checking and verification based on data from these sources needs to be carried out by journalists using various specific practices like, for example, reverse-image search, geolocation, and web scraping (Brandtzaeg et al., 2016). At a more basic level, editors also need to acquire training and knowledge in terms of data privacy regulations (Reventlow, 2020) and the ethical use of data in journalism practice. This includes considerations on automated news personalization and the potential risks associated with it (Zuiderveen Borgesius et al., 2016). Overall, data literacy has become an increasingly critical skill in journalism and beyond (Carmi et al., 2020), thus forming an essential part of journalism training and education (Gray, Bounegru, & Venturini, 2012).

At the same time, the gatekeeping function held by news media has been partly relinquished to non-journalistic entities like social platforms and aggregators (Coddington, 2020; Wallace, 2018). Add to this the ongoing transition from a "more or less coherent industry to a highly varied and diverse range of practices" (Deuze & Witschge, 2017, p. 167) spurred on by various factors that include an always-online mode of "networked production" (Van Der Haak et al., 2012), influencer journalism or micro-blogging (Maares & Hanusch, 2020; Holton, Coddington & de Zúñiga, 2013)—digital technologies enabling anyone

to potentially reach a global audience, which in itself challenges traditional journalistic role perceptions (Newman, 2018).

Such circumstances lead to mounting pressure on editorial staff to understand and apply new media technologies (Schätz & Kirchhoff, 2020, pp. 104–105) and to be able to evaluate and push for innovations in the field. It is against this backdrop of professional uncertainty paired with constant innovation pressures, changing media repertoires, and tooling that the following phenomena should be considered.

3.5.1 Editorial analytics

As editorial analytics have played an integral part in the now-diminishing business model of display advertising[39], research on the topic is plentiful (Tandoc, 2019; Cherubini & Nielsen, 2016). But how have editorial analytics evolved and how do they contrast with other types of analytics? On a more abstract level, analytics software could be said to enable operators of digital applications or machines to collect and analyze operational or usage data in an effort to optimize said applications for arbitrary goals. More specifically, analytics are often thought of in terms of audience characteristics and detailed statistics around the transactions happening on any given website or networked mobile application (Tandoc, 2014; Kaushik, 2009). In the field of healthcare, analytics are used to identify disease trends via predictive modeling, improve patient care or guide medical workers towards more accurate diagnoses and treatment plans (Raghupathi & Raghupathi, 2017). Telecommunication companies like phone or internet providers might use analytics to detect fraud, predict customer churn (Jadhav & Pawar, 2011) or manage their general network performance (Zahid et al., 2019). In finance, analytics similarly help with credit risk analysis (Baesens et al., 2016) or predict credit card fraud (Broby, 2022). In essence, analytics involves the systematic computational interpretation or analysis on the basis of data or statistics—which in turn requires technological affordances that integrate data storage, data aggregation, and interfaces to query and visualize the data.

[39] The advertising landscape around online journalism is complex, with some sections growing while others decline, but overall, publishers have shifted their reliance on advertising revenue (as a function of reach) to other types of digital business models such as subscriptions (Newman, Fletcher et al., 2022, p. 18; Chyi & Tenenboim, 2016). As an example, the display advertising revenue of the New York Times continues to decline year-over-year (Guaglione, 2024).

In order to establish an even clearer understanding of the fuzzy language around analytics, a distinction could be made between different sub-categories of analytics that are prevalent in the field. As one such subcategory also linked to the emergence of the internet of things (IoC), *telemetry* is a more specific type of analytics that involves the automatic measurement and transmission of data gathered from remote or inaccessible sources in physical space, often through sensors which communicate through wireless means.[40] Such data might include animal movements (Hussey et al., 2015), atmospheric pressure (Li et al., 2009), or vehicular traffic flow (Nguyen, Dow & Wang, 2018). Telemetry is often used extensively in the field of space exploration, where sensors gather data from satellites and other unmanned spacecraft (Zhan et al., 2020).

Another concept borne out of the software industry, business intelligence (BI) could be said to combine data mining, data visualization, infrastructure and data practices into a particularly marketable package.[41] While analytics, BI and concepts such as decision making systems (DSS) are "traced and interwoven as they appear to converge and diverge over the years" (Power, 2007, p. 1), BI more specifically refers to the tools, software and systems that aid in the decision-making processes of businesses. Another adjacent concept, *predictive analytics* uses statistical algorithms and machine learning techniques to identify the likelihood of future outcomes—the field upholds the idea of obscure patterns, new correlations, market trends, customer preferences, and other business information waiting to be uncovered (Eckerson, 2007; Larose, 2015; Kumar & Garg, 2018).[42]

---

[40] An adjacent type of journalism would be "sensor journalism" (see also Carlson, 2015; Loosen et al., 2022), which in this sense could also be thought of as "telemetry journalism".
[41] See also p. 50; 8.1 "Business Intelligence (BI)"
[42] See also 8.1, "Predictive Analytics/Predictive Learning"

Narrowing our focus again to analytics as applied in the field of journalism, here analytics are used to gain insights into audiences, track the performance of content and make data-driven decisions on the basis of varying and changing metrics[43] such as pageviews, unique users, time spent, bounce rate or traffic by acquisition channel (chronologically sorted; Cherubini & Nielsen, 2016, p. 34). In this study, I will refer to analytics practices and software used by editorial staff as editorial analytics—with the expectation of other forms of analytics or analytical data work happening *outside* of the editorial domain.

In the late 2000s, analytics companies such as Chartbeat or Parse.ly gained traction in newsrooms, targeting online publishers specifically and aligning their value proposition with journalistic values (Belair-Gagnon & Holton, 2018; Petre, 2020).[44] In the seemingly "dispassionate dashboards" (Petre, 2015, p. 24) provided by these analytics providers, audience data indicates to the editorial staff how their work performs according to a multitude of criteria. Scholars have subsequently studied the ways in which editorial analytics have been adopted as metrics of success for content and audience engagement (Bunce, 2019; Duffy, Ling & Tandoc, 2018), how they have forced journalists to rethink their professional processes (Tandoc & Thomas, 2015), and how journalists work within a newsroom culture that places, in some cases, more value on analytics than their professional intuition (Hanusch, 2017). Even though newsrooms across the world have incorporated editorial analytics into their daily practices, scholars are hesitant to suggest that journalists should adopt audience-driven data into news judgment practices (Anderson, 2011; Nguyen, 2013; Tandoc, 2014; Zamith, 2018).

[43] A glossary entry for the term metrics can be found in chapter 8.1, "Metrics".
[44] Google Analytics was first made available in 2005; Chartbeat and Parse.ly were both founded in 2009 (Crunchbase.com, last accessed Aug. 12, 2023)

By allowing a (not quite) real-time observation of the audience, researchers find, analytics methods can lead to a thematic narrowing of the news offering and a "self-reinforcement of mass taste" (Neuberger & Nuernbergk, 2015, p. 200, translated by the author) in the long term. More recent studies point to the metrics discourse as a standard in determining the epistemic value of news, manifesting in newsroom strategies, guidelines, and discussions. Such metrics practices are encouraged through coaching, evaluating and rewarding individual journalists' performance (Ekström, Ramsälv & Westlund, 2022, p. 755). Additionally, metrics are actively reconciled with journalism's independent standards, emphasizing the provision of relevant and verified public knowledge about current events. Furthermore, a study reveals how the embrace of metrics "radicalizes the focus on presentation, packaging, and timing in the optimizing of material" (Ekström, Ramsälv & Westlund, 2022, pp. 767–768). Asked directly, journalists recognize the value of audience data for news organizations and acknowledge the larger cultural dimension of audience data practices, such as relying on quantitative metrics for editorial decision-making—while often not claiming responsibility in their analysis (Schaetz, 2023).

On the other hand, audience analytics can generate positive effects when they go beyond mere numbers, functioning as "participatory mechanisms" that "depict human behavior" (Blanchett, 2021, p. 14). As a result of technological advancements, audiences can now actively participate in editorial decision-making both pre- and post-publication (Blanchett, 2021). Such data gathered directly by publishers and given explicitly by users, also known as *first-party data*, gains relevance with the imminent demise of third-party cookies, however, news websites are still facing a severe trust issue as consumers are reluctant to share personal data with them (Newman, Fletcher et al., 2022, p. 11).

Overall, as the increasing relevance of analytics data in journalism remains among the more investigated topics in journalism research, researchers are continuing to monitor "metrics-driven journalism" (Cherubini & Nielsen, 2016; Zamith, 2018) in which editorial decisions might strongly point towards reach or other arbitrary quantitative goals.

Although I argued that the relevance of editorial analytics has diminished over time, many challenges and questions posed by scholars regarding analytics remain timely and warrant further re-examination. Are dashboards still talked about as if they were "dispassionate" (Petre, 2015, p. 24)? Are web or editorial analytics tools and metrics still operated by and pushed into the field by influential outside companies? What level of responsibility or authority over analytics systems do data workers acknowledge? During interviews, another meta-discussion might also evolve around the implicit (collected without direct user input) versus explicit (actively entered or given by users) nature of analytics data.

3.5.2 Data as a means of reporting

In short, data journalism or data-driven reporting refers to data as a basis for or means of reporting (Tong, 2022) and should not be confused with journalism that is *about* data.[45] Often cited as another driver of journalism's quantitative turn (Coddington, 2015), data journalism as a practice involves the collection, analysis and visualization of large datasets to uncover patterns or insights that are then incorporated into news stories or interactive visualizations (Gray, Bounegru, & Chambers, 2012). Employing statistical methods, computational tools and data visualization, a data journalist might look to either enhance the depth and context of broader news stories or deliver data as the reporting artifact itself (Howard, 2014; Parasie & Dagiral, 2013). After first emerging in the 2010s, data journalism re-gained significance during the COVID-19 pandemic (Desai, Nouvellet et al., 2021) and to this day a considerable number of prizes are awarded to data-related journalism (Schätz & Pühringer, 2022). In the following, I want to discuss how data journalism introduced more universal challenges of data work into the editorial day-to-day, look at specific examples, and touch on the valid concerns that have been raised around this type of journalism.

First, data journalism requires data and the awareness of (publicly) available quantitative data as a source. Historically linked to the phenomenon, the open data movement has had a significant impact by introducing its ideas and norms into the field: transparency, accountability and innovation by making data freely available to citizens, researchers and businesses (Janssen, Charalabidis, & Zuiderwijk, 2012).

[45] Data-related news topics would include data privacy, artificial intelligence, disinformation, big data and surveillance or more recently, how governments introduce data into their decision-making in crisis mode with the COVID pandemic.

The movement has led to a shift in journalistic practices, emphasizing the importance of open data sources and the use of digital tools for data analysis (Baack, 2015) and creating an entanglement between journalists, activists and civic technologists around data-driven stories (Baack, 2018). A prime example of such a collaborative effort between activists, technologists and journalists, the *Panama Papers* demonstrated the significance of specific data practices (Heft, 2019; Baack, 2016).[46] Members of the International Consortium of Investigative Journalists (ICIJ) used advanced methods for cleaning, ordering, and presenting the data. While the *Panama Papers* team did not employ machine learning (ML) methods, one example of a replicable editorial production process using ML can be found at the Washington Post Computational Journalism Lab (Schmidt, 2019).[47] In the context of a data-driven story on the Democratic presidential candidates, reporters collected thousands of tweets and mapped these to political issues using a technique called clustering. An exemplary application of the "human-in-the-loop" principle, which integrates machine learning with human oversight and manual adjustment, can be observed in the practices of the Washington Post (Fails & Olsen, 2003). Overall, we can establish how data journalism introduces highly sophisticated technical skills into the editorial production cycle, fostering a professionalization around data and computational methods in the field while also encouraging cross-border journalism (Alfter, 2016) between multiple public spheres to different degrees (Loosen, Reimer & De Silva-Schmidt, 2020).

[46] While not primarily quantitative in nature, the cross-functional team here needed to extract meaning out of massive amounts of information using techniques like pattern recognition and extract, transform and load (Cabra & Kissane, 2016). See also 8.1, "Extract, Transform and Load".

[47] According to director at the ICIJ at the time, machine learning methods would significantly accelerate the process in future works of similar scope (Guevara, 2019).

This kind of professionalization has its limits, as a study around data journalism education finds journalists interested in data are predominantly highly educated in journalism or closely related fields, yet not having a strong level of training in the more technical aspects of data journalism such as data analysis, coding and visualization (Heravi, 2019). Additionally, even though specific consecutive or post-graduate study programs for data journalism do exist, the field overall does not have a strong academic underpinning (Heravi, 2019).[48] Depending on the geographical frame of reference, data journalism might be restricted by the availability and accessibility of public datasets (Loosen, Reimer & De Silva-Schmidt, 2020). In a recent study, reliance on government data sources hints at heavy institutional influence at legacy news outlets, whereas non-legacy organizations were found to rely less on government data and lean more towards self-collected data (Lowrey & Hou, 2021). However, as other studies find, data journalists are developing strategies to counteract their dependency on major data sources and even collect their own data, for instance through collaborations and networks (Porlezza & Splendore, 2019).

Echoing the theoretical premises of this study—data as always cooked to some recipe, never neutral but rather part of a larger data regime—some contend that objectivity in data journalism is a fallacy, with data reporting a process of knowledge construction, determined by factors like the use of specific algorithms in the preparation of data, hurdles around verification, the potentially insufficient understanding of data contexts by reporters and what Tong and Zuo refer to as "design subjectivity" (Tong & Zuo, 2021, p. 2).

[48] As of the time of writing, the Master of Science in Data Journalism at the Columbia Journalism School or several PhD graduate programs from e.g. the University of Sheffield. A dedicated Public Affairs Data Journalism course at Stanford University seems to have phased out in 2018.

For these reasons, scholars propose a shift in academic attention from the celebration of objectivity to the study of the epistemology of data journalists (Tong & Zuo, 2021). On the level of content matter, data journalism also introduces some new issues. As one study concludes, this type of journalism leads to an increase in abstract constructs alongside a decrease in anecdotal knowledge, highlighting a demand for the constant monitoring of statistical practices in journalism to counter misuse and foster awareness of these increasingly abstract constructs (Lowrey & Hou, 2021). Another specific issue of data as an output of reporting lies in the presentation of data through graphical interfaces or interactive applications. As opposed to text-based articles, the preservation, archival and searchability of data journalism becomes less straight-forward. Traditional news archiving does not yet have systems in place for preserving these outputs (Heravi et al., 2022).[49]

Condensing the literature reviewed here, while data journalism clearly caused some level of professionalization around data and data practices in the field, I do not find strong evidence of entirely new classes of workers stepping in; neither do structural shifts outside of the editorial sphere appear to take place as a consequence of the phenomenon. Instead, scholars point out how *more* data literacy and competence would be required to fully develop and stabilize data journalism's potential.

[49] Ideas range from keeping a working version of visualizations available via methods such as emulation, migration or virtual machines (VMs); or attempting to capture a flat or simplified version ("surrogates") of these applications (Heravi et al., 2022, p. 2094)

3.5.3 Data around the newsroom

Having covered the larger data-adjacent research themes of (editorial) analytics and data journalism, I want to briefly discuss other relevant areas where data plays a role around the newsroom.

As mentioned in previous chapters, the expectations around skillsets and the abilities of journalists are changing as new practices such as analytics and data-informed reporting emerge. While journalists are increasingly expected to have "basic statistical knowledge and operate Excel" (Schätz & Pühringer, 2022, p. 11, translated by the author), the ongoing professionalization of the field also leads to entirely new work profiles. As a key technological role inside the newsroom, editorial technologists are individuals with programming and computational skills who work at the intersection of journalism and technology—developing news bots, text automation, storytelling visualizations or news recommender systems (Lischka, Schaetz & Oltersdorf, 2022). Partly engaging in data work, they accumulate symbolic, cultural and social capital in news organizations, possess collective agency for change and hold several editorial-technological doxa around algorithmic designs or their responsibilities towards certain parts of the organization (Lischka, Schaetz & Oltersdorf, 2022). Consequently, they strive for recognition within editorial offices and become integral parts of multi-skilled editorial teams as they are clearly entrenched in the editorial domain, struggling with being seen as magical or "special unicorns" due to the uniqueness of their skills while attempting to demystify their abilities and make them accessible to everyone (Lischka, Schaetz & Oltersdorf, 2022, p. 1033). Interestingly, editors estimate the structural changes brought about by technological specialists to be relatively modest (Schätz & Pühringer, 2022, pp. 19–20).

During my interviews the debate around a) data literacy, what journalists should be able to accomplish with data of their own accord and b) inter-organizational responsibilities will most likely continue.

In recent years, digital journalism has seen a surge in advanced computational methods like machine learning (ML) techniques to enhance news production, distribution and consumption (Diakopoulos, 2019). Multiple substantive concepts and technologies have emerged or evolved in the context of digital journalism that rely on data to work properly: data mining, recommendation and personalization.[50] Data mining is the process of discovering patterns, relationships, and insights in large datasets through the use of various computational techniques and algorithms (Han, Kamber & Pei, 2012), which can be particularly helpful in the research and exploration phases of news production. Several data mining methods have been identified as particularly relevant to news production (Diakopoulos, 2019, p. 43):

- Clustering of similar objects
- Object classification
- Regression analysis
- Automated summarization
- Modeling of dependencies
- Determination of deviations

[50] Both quality and quantity of training data directly impact the performance of ML models, which emphasizes the importance of collecting and maintaining comprehensive, diverse and representative data sources (Halevy, Norvig & Pereira, 2009). Generally, without data, ML would be unthinkable (Gröger, 2021).

Here, classification corresponds to so called supervised learning, a form of machine learning designed to prepare and label data. Objects are classified according to previously defined rules. In the process of unsupervised learning, on the other hand, the model operates without predefined classes, which must first be determined by the machine, for example by means of clustering within a data set. Similarities are determined statistically and then expressed as confidence values. The qualification and assignment of meaning to what makes up a cluster remains an editorial task. It remains to be seen if, and to what extent, such advanced methods will be addressed by participants in the present study.

Personalization is another technological advancement often discussed and increasingly investigated in the field of digital journalism. News organizations leverage software algorithms to create personalized news feeds, offering users a curated selection of stories that cater to their preferences and interests (Thurman & Schifferes, 2012). By leaving various digital traces through interaction with digital news affordances, users also become quantifiable targets of various metric-driven strategies. Originally a matter of increasing advertisement revenue, news personalization shifts to targeting users with personalized content as publishers move from an ad-financed business model to subscriptions or paid news (Bodó, 2021). Machine learning techniques such as text analysis and topic modeling are employed to categorize content and match it with segmented user groups (Loosen & Solbach, 2020).

A subset of personalization, news recommender systems (NRS) analyze user behavior (implicit data), as well as preferences and interests (explicit data) to suggest relevant content to individual users (Karimi, Jannach & Jugovac, 2018)—with vastly different results depending on the underlying mechanisms.[51] From the publisher perspective, recommendations are employed to drive users towards becoming paying subscribers.[52] Researchers have found NRS to mitigate information overload by providing users with tailored news content based on their preferences and interests, leading to a more satisfying and efficient news consumption experience (Liu, Dolan & Pedersen, 2018). Moreover, studies have shown that NRS can foster diversity in news consumption by exposing users to a wider range of topics and perspectives than they would typically encounter through manual browsing (Helberger, Karppinen, & D'Acunto, 2018). With the arrival of large language models (LLMs) and generative AI, news personalization might even happen, at the level of content, intratextually.[53] But such fully automated scenarios of autonomous agents doing editorial work are unlikely to happen inside of news organizations, as studies found "audiences tend to make societally suboptimal choices about what news or information they consume, especially if technologies are deployed to exploit human weakness in order to turn a profit" (Bodó, 2021, p. 1071).

---

[51] Including personalized and content-aware (topic modelling, item-based, keyword based, sequential pattern mining), personalized and user-based (collaborative filtering, bayesian personalized ranking, k-nearest neighbor) and non-personalized (item-2-item, recently popular, trending) recommendations (Karimi, Jannach & Jugovac, 2018).

[52] "What are the hooks for us that are going to bring people in and make them commit to us, engage with us on a regular basis and ultimately become subscribers? Recommendation is the obvious answer." (Rockwell, 2019).

[53] An idea that precedes LLMs and generative AI, advances in automatic text generation would allow for content personalization at the article level, adapting style and text elements to user preferences per request (Loosen & Solbach, 2020, p. 192). A combination of personalized content *and* recommendations raises fundamental questions about the social representation of news and the balance between individual preferences and shared sources of information.

As mentioned above, several of the stimuli (around data literacy, conflicting responsibilities of technologists, advanced computational methods, implicit data collection, or personalization) collected in this chapter could either be integrated into my line of questioning or re-appear naturally in conversation with the participants of the study.

4. Objectives

4.1 Overview

After establishing a theoretical framework that matches my overall research interests, I have reviewed relevant literature and concepts as well as prior research in journalism studies and beyond in the previous chapter. Among other things, I have acknowledged how data work is shaped by the "profoundly ideological role" (Van Dijck, 2014, p. 5) of individual beliefs in its power. I have also identified datafication as a remarkably modernist narrative and mode of operation, in some ways a continuation of the industrious organization around machines. On the organizational level, which this study is mainly interested in, datafication might turn out to be a less straightforward or even reversible process. These are considerations that begin to point towards specific questions. Would actors at the organizations under investigation consider current data work as a continuation or revolution? Will I encounter a certain ideology around data? In this chapter, I attempt to condense these preliminary discussions and reflections on data, data work, and the theoretical framework into a number of guiding assumptions, which then lead me to my research questions. As there is overlap between the layers of theory, some questions will feed into multiple sections during synthesis. Similar to the ways in which the question of isomorphic qualities spans all organizations, the "provocations" of critical data studies become a cross-cutting concern. As we have gathered from CDS, data should be thought of in the context of wider data assemblages and data regimes "doing active work in the world" (Chapter 2.2.1). An overarching goal, then, would be to uncover or deconstruct both the sociotechnical data assemblages I encounter in the field as well as the structures of power that might have established said assemblages or replicate themselves through them.

4.2 Research questions

As explained in the previous subchapter, I consider the "provocations" posited by critical data studies (CDS) as a cross-cutting concern and will discuss these as a synthesis across cases. Descending downwards from CDS in my theoretical framework, organizational isomorphism (Chapter 2.2.2) describes the tendency of organizations to adopt similar structures, practices, and processes as their competitors in the same industry. Within the context of news organizations, I can formulate the assumption that an increasing importance of data in the media industry has prompted a convergence of data-related practices and structures among news organizations. This would be evidenced by observable structural changes, such as the creation of new positions or departments dedicated to data management, analysis, and application. It is likely, therefore, that the changing significance of data in the media industry has led to a more pronounced isomorphism across news organizations as they strive to adapt to shifting industry trends and practices. By examining data-related practices, job titles, and reporting structures across multiple news organizations, insights into the factors that drive structural change in response to data's increasing importance in the industry can be gained.

Guiding Assumption

Ways of working with data in news organizations have fundamentally changed in recent years and in turn these organizations underwent structural changes to reflect and/or act upon these changes.

Research Question RQ1

How have news organizations shifted or enhanced their ways of working with data in recent years and can we identify patterns at an inter-organizational level?

In order to allow for a comprehensive look at data assemblages (Chapter 2.2.1), I extend my line of questioning towards technical details and broaden our knowledge about the contexts of data work. As a reminder, data assemblages are thought of as made up of two parts, with the first one being the technical stack, "instrumental means by which data are generated, processed, stored, shared, analyzed and experienced" (Kitchin, 2022, p. 23).

Guiding Assumption

Data work at the organizations under investigation is governed by specific infrastructure, software and data affordances.

Research Question RQ2

How are data generated, processed, stored, shared, and analyzed and what can we learn about the specific infrastructure, software and data affordances used to facilitate these activities?

Alongside the technical stack, "a number of discursive and material components related to philosophy and knowledge, finance and politics, law and governance, practices, stakeholders and actors, geography and markets" (Kitchin, 2022, p. 25) are thought of as constituting the contextual stack. While the technical stack appears to be an inherent part of data assemblages or regimes, the contextual stack becomes so vast in its scope that I need to limit myself here to the immediate organizational context. Again, these contextual conditions of data work appear so fuzzy that I cannot formulate a specific research question without considerable redundancy to other questions that address, for example, the normative pressures or market forces that might lead to organizational isomorphism. To approximate complete data assemblages I instead attempt to (non-exhaustively) map apparatuses (systems of thought, forms of knowledge,

practices, institutions, places) and their respective elements that appear in the material overall. However, it is unknown how these new instruments and apparatuses represent something different from previous ways of working with data. To understand this, we need to ask about historic developments at the news organizations being studied. Crucially, while digital affordances have facilitated many new data-related practices, it is essential to explore whether or not these have reshaped the fundamental nature of data work altogether.

Guiding Assumption

Current data work at news organizations represents a new quality (in terms of professionalism, investment, and/or volume) as compared to previous iterations of data work.

Research Question RQ3

How does data work in its current form differ from previous ways of working with data, given the assumption that data work does not require digital affordances per se?

Looking at the three mechanisms of isomorphic change posited by DiMaggio & Powell, determining mimetic processes (DiMaggio & Powell, 1983, p. 151) empirically appears to be rather problematic for the scope of this study, as we can expect interviewees to deny any and all imitation of competitors to achieve legitimacy, acquire resources or otherwise cope with an environment of uncertainty. Rather, I want to probe for normative and coercive forces at play and then discuss mimetic isomorphism through a process of interpretation and exclusion. In the context of news organizations, coercive pressures to acquire funding sources, adapt to societal expectations or incorporate regulatory requirements might very well lead to a homogenization of behaviors and

practices related to data work. Following the guiding assumption of coercive forces at play, we can improve our understanding of how organizations navigate around these forces.

Guiding Assumption

There are external factors which compel news organizations to behave uniformly and synchronously when it comes to the adoption and execution of data work.

Research Question RQ4

Why have news organizations adapted or enhanced their ways of working with data in recent years and can we find similarities in their origins and reasoning?

Isomorphic change can also be a consequence of the normative pressures of a professional class "establishing a cognitive base and legitimation for their occupational autonomy" (DiMaggio & Powell, 1983, p. 152). This makes it important to examine the professional backgrounds and normative influences of individuals in data roles. One possible finding might be that data workers in our sample appear non-diverse. This could be due to a number of factors, such as the prevalence of certain degree programs, experience in particular industries, shared career paths, or data analysis backgrounds. If this were the case, a lack of diversity in these backgrounds could potentially limit the range of perspectives and expertise represented in news organizations.

Guiding Assumption

The professional backgrounds of data workers introduced to the field in recent years are non-diverse.

Research Question RQ5

What kinds of professional backgrounds (and normative influences) do data workers at the examined organizations have?

The shift towards data-driven media practices has the potential to affect the internal dynamics of news organizations. Specifically, data practices and data affordances may either reinforce pre-existing interdepartmental boundaries or create new ones. One may argue that data work could also provide a means to transcend these boundaries by facilitating collaboration and collective problem-solving across departments through boundary objects (Chapter 2.2.3). However, others may counter that data work entrenches these boundaries through the specialization and centralization of data management.

Guiding Assumption

Data work and data affordances, while facilitating collaboration to some extent, at the same time reinforce interdepartmental boundaries due to the specialized and centralized nature of data and its management.

Research Question RQ6

How would data affordances and objects work towards either transcending or reinforcing inter-departmental boundaries between the editorial and publishing domain or other demarcations?

David Beer's concept of the data gaze (Chapter 2.2.4) refers to the shifts in power relations that accompany the use of data analysis and interpretation in a variety of fields, including journalism. The expertise required for interpreting data-driven information, however, is not solely limited to data journalists. This might be reflected in the fact that various data roles have emerged on the publishing side, sharing some amount of interpretative power. These individuals are tasked with analyzing complex data sets, making sense of the information presented, and drawing meaningful conclusions from it. As such, they play a significant role in shaping the way that data is used by news organizations more broadly. The notion of expertise as "interpretative knowledge" (Chapter 5.2.3) is particularly relevant in this context, as it highlights the significance of not only analyzing and interpreting data, but also communicating it in a way that should be accessible and comprehensible to non-experts. With these ideas in mind, I assume that the emergence of these data roles have led to a re-distribution of interpretative power in news organizations.

Guiding Assumption

Next to data journalists and audience analysts who more or less shares the same ethical and professional standards of journalistic production, new data-related roles with deviating norms and backgrounds have emerged on the publishing side which exert a great deal of interpretative power.

Research Question RQ7

How have new data-related roles emerged within the publishing side of news organizations and to what extent do these roles, with potentially differing norms and backgrounds, influence the interpretation and application of data in journalistic production?

Continuing further with the notion of the data gaze (Chapter 2.2.4) and individual perspectives of data workers, we need to question how individual interviewees reflect on the purposes and agenda-setting power of their own work. This exploration could reveal insights into the motivations that drive data interpretation and reporting, as well as any ethical considerations or conflicts that may arise.

Guiding Assumption

Inside news organizations, specialist knowledge workers tasked with the analysis and interpretation of data wield significant subjective power, intentional or non-intentional.

Research Question RQ8

How do data workers reflect on their own agency, potential conflicts of interest and the agenda setting power of their own work?

As demonstrated, research on data work in the sense of an emergent professional domain, the byproduct of general digitalization, can be predominantly found in a handful of disciplines like information studies, health sciences, computer science and economics (Chapter 3.4). While not tied to any specific theory or assumption, I will address the question of what constitutes data work in journalism—with the goal of formulating a broader definition of the term within the field.

Guiding Assumption

Data work in journalism, despite variations among news organizations and teams, can be distilled into a generalized definition of shared qualities and properties.

Research Question RQ9

How can we conceptualize the qualities and properties of data work across all actors and cases?

5. Research

5.1 Overview

The empirical basis of this study consists of 21 qualitative interviews (n=20, one test interview) conducted specifically for the thesis by the author, spread across 6 case studies inside both established ("legacy") news organizations and digital-native news organizations. In total, with an average duration of 1.05 hours, 22 hours of material were analyzed. Why case studies? Examining types of isomorphisms in a field requires comparing multiple organizations. Additionally, as we have seen, critical data studies establishes data as a phenomenon of high complexity, with technological assemblages and data worker self-perceptions possibly varying greatly between organizations. These research subject properties suggest a multiple case study design.

5.2 Research design

5.2.1 Case studies

Case studies emerged from their original application in medical and psychosocial research (Creswell & Poth, 2018, p. 154) to attain broader adoption and have sustained their popularity as a research method across various disciplines (Yin, 2018, p. 17; George & Bennet, 2005, p. 26). For instance, case studies are widely used in a variety of social sciences (Robinson, Acemoglu et al., 2012; Greenwood et al., 2011; Edmondson & McManus, 2007), including media studies (Lewis & Usher, 2016; Diakopoulos, 2014). They are often employed in situations where the aim is to explain presumed causal relationships that are challenging to map with surveys or experiments (Yin, 2018, p. 19). For studies about organizations and phenomena within their real-life context and outside of the controlled conditions typical of experiments, case studies are considered a suitable choice (Eisenhardt & Graebner, 2007, p. 25). Particulary, case studies are valued for their ability to explore a limited set of research questions starting with "how" or "why" (Yin, 2018, p. 44) and for their capacity to develop and refine theoretical constructs for future research (Baxter & Jack, 2008, p. 544). Although often used to generate hypotheses (Creswell & Poth, 2018, p. 146; Eisenhardt, 1989, p. 546), case studies can also be descriptive or exploratory in nature (Yin, 2018, p. 8; Flyvbjerg, 2011, p. 306). Striking a balance between both ends, I adopt an exploratory case study approach to investigate research questions informed by a set of pre-existing theoretical building blocks (Bryman, 2016, p. 65). I also understand case study research to inform my overarching research design while data collection and data analysis methods are selected separately based on my particular research objectives (Creswell & Poth, 2018, p. 153).

The two prominent methodologists Robert Stake and Robert Yin represent two distinct approaches to case study research. Yin could be seen as rooted in a post-positivist tradition, emphasizing methodological rigor, construct validity and reliability, advocating for the use of protocol and systematic procedures to ensure replicability and generalizability of findings to broader theoretical propositions, a process he refers to as analytic generalization (Yin, 2018, p. 79). On the other hand, Stake adopts a more constructivist perspective, emphasizing the inherent uniqueness of each case (Stake, 2010, pp. 31–32). His approach focuses on the intrinsic interest of the case, the "special something" (Stake, 1995, p. 133), rather than its potential for generalization. Stake argues that the primary concern should be to increase understanding and extrapolate lessons from the case itself, rather than to generalize findings alongside other cases (Stake, 2010, p. 182; Stake, 1995, pp. 7–9). Overall, Yin's systematic and structured approach aligns more closely with the deductive nature of my research design. A basic distinction in research design can be made between single case studies and multiple case studies (according to the number and structure of the cases). Yin further distinguishes between a total of four case study variants (Yin, 2018, pp. 97–107) according to the number of cases and the units of analysis inside the individual case:

- *single-case holistic designs* (type 1)
  Single case, no differentiation in investigation units in the case.
- *single-case embedded designs* (type 2)
  Single case, multiple units of inquiry inside
- *multiple-case holistic designs* (type 3)
  Multiple cases, no differentiation in units of inquiry per case and
- *multiple-case embedded designs* (type 4)
  Multiple cases, multiple units of inquiry across cases

If funding and time permits, a multiple-case study design should be preferred over a single-case study (Yin, 2018, pp. 24–25). The obvious advantage of multiple-case studies lies in their replication logic and a capacity to produce generalizable patterns, ideally contributing to an increased testability of theoretical findings (Yin, 2018, pp. 102–103; Creswell & Poth, 2018). In a holistic case study, the researcher looks at the case in its entirety, considering the overall context and treating it as a single unit of analysis (Yin, 2018, p. 50). This method appears particularly advisable where a complex, multifaceted phenomenon can not be easily divided into individual components (Creswell & Poth, 2018, p. 74). When deciding between a holistic and an embedded case study, it is important to consider that holistic approaches are especially suitable when no meaningful units of investigation can be formed, or when the underlying theory emphasizes a holistic context. Since my focus lies on the isomorphisms of data work shared among multiple organizations, the components of the theoretical framework in this thesis are predominantly aimed at the organizational level. Therefore, a holistic multiple-case study design (type 3) is most appropriate.

Targets of investigation in case studies are usually one or more phenomena, organizations, industries or even policies (Stake, 2010, pp. 25–26). A wide variety of sources and methods can be considered for data collection in case studies. Yin distinguishes between six sources of evidence including documentation, archival records, interviews, direct or participatory observations, and even physical artifacts, albeit to a lesser extent (Yin, 2018, pp. 178–193).

Documentation might include, for example, personal emails, internal company documents, studies, or press articles, and an increasing number of potentially interesting documents are now publicly available via the internet (Yin, 2018, p. 179). Although internet documents in particular may not always be accurate in terms of content and pose the challenge of overabundance, they can be used to supplement other sources (Stake, 2010, p. 116) after consideration and triage as to their centrality to the individual research interest (Yin, 2018, p. 181). Although one data source is sufficient for data collection and many studies rely on one source alone, it is beneficial to use multiple sources (Yin, 2018, pp. 114–115).

Qualitative methods such as interviews are among the most popular approaches to obtaining data in case study research (Eisenhardt & Graebner, 2007, p. 28; Yin, 2018, p. 183; Benbasat, Goldstein & Mead, 1987, p. 381) and qualitative-empirical methods are generally considered to support the study of complex phenomena in great depth (Creswell & Poth, 2018, p. 150). This study relies primarily on interviews, making use of secondary sources of information in the form of documents, such as blog posts, public-relations content disseminated by the organizations, or documentation on the data technologies and affordances mentioned by the interviewees. Furthermore, to corroborate and cross-reference statements in connection with emerging data roles and general corporate structure, archives of job listings are considered. In the following sections, the expert interview will be discussed in detail as the dominant data collection method used in this study.

5.2.2 Sampling criteria

As I have established, a holistic multiple-case study design offers a robust framework for the exploration of complex phenomena within organizational contexts. As the next step, the selection of cases and interviewees can significantly impact the validity and reliability of the findings. Here I discuss potential guidelines to aid the selection.

First, the criteria for selecting the cases themselves should be based on the purpose of the research and the nature of the phenomenon under investigation. As I am studying phenomena at the organizational level, I follow a case (or "site") selection path around characteristics such as industry, organizational structure or technologies (Benbasat, Goldstein & Mead, 1987, p. 373). In the context of media organizations, the set of criteria might include factors such as the size of the organization, its geographical location, the type of media it produces, its market share or target audience (Lindlof & Taylor, 2017, p. 122). Exploratory case selection often requires the researcher to rely on externally ascertainable characteristics (Yin, 2018, pp. 68). In most situations, the sampling would aim for replication across multiple cases that appear typical or representative of these characteristics (Seawright & Gerring, 2008, pp. 299–300). Deviant cases, which by reference to some general cross-case relationship, demonstrate a surprising value, might also be used to great effect in the context of purely exploratory studies (Seawright & Gerring, 2008, p. 301). To obtain as much information as possible about a particular phenomenon, the selection of an extreme case could even be more suitable than a representative case (Flyvbjerg, 2006, pp. 29–30). Overall, ideal case selections provide the opportunity to maximize what can be learned (Stake, 1995, p. 4) while still following a replication logic rather than a random sampling logic (Yin, 2018, pp. 91–93; Eisenhardt, 1989, p. 542).

As for the ideal quantity of cases, multiple case study investigations are feasible with as few as two replications, but ultimately this depends on the number of desired literal and theoretical replications (Yin, 2018, p. 94). In multi-case research design, the number of cases typically ranges from four to ten, as this allows for a balance between achieving depth and enabling meaningful cross-case analysis (Eisenhardt, 1989, p. 545; Yin, 2018, p. 105). This study limits the number to six cases covering a diverse spectrum of geographical locations (spanning all of Germany), organization sizes (from small to large enterprises) and distributions (digital-native, regional, and national) while possessing shared traits in the type of media they produce (digital journalism) and operating in the same domestic market (Germany).

A selection of interviewees within these case studies should also be carefully considered to ensure a diverse range of perspectives. Criteria for selecting interviewees might include their role within the organization, their level of experience in the media industry, their area of expertise, and their willingness to participate in the study. Research into potentially interesting candidates was carried out in advance of approaching the individual via email and business networks. I opted to engage mostly with executive-level personnel for their abilities to authorize participation on their own behalf (Benbasat et al., 1987, p. 373) and to provide the maximum amount of information due to their status as gatekeepers and "savvy social actors" (Lindlof & Taylor, 2017, pp. 177–178). Another tendency I need to acknowledge, is that I selected contacts based on their assumed influence on data-related topics inside each organization (Yin, 2018, p. 69). Additional interview partners were then acquired either through these initial contacts associated with the case or by way of referrals through snowball sampling (Atkinson & Flint, 2001).

To prevent systematic distortions in the impressions gathered during interviews, several countermeasures can be implemented, such as selecting experienced interview partners from diverse areas, hierarchical levels, or locations to ensure a variety of perspectives (Eisenhardt & Graebner, 2007). These considerations determine the number of interviews required for any given study while especially in the case of doctoral dissertations, several constraints around timing and availability of interviewees might lead to compromise. In terms of the average sample size in qualitative disserations, the literature produces an ambivalent picture.[54] As a larger number of interviews does not necessarily lead to more useful information (Mason, 2010), representativeness considerations then become of secondary importance in interview-based studies while "information power" (Malterud, Siersma & Guassora, 2015) takes precedence. Instead, I adopt the concept of saturation (Glaser & Strauss, 1999) to determine the rational cut-off point. Additional interviews were conducted until additional data no longer yielded new insights and saturation was reached with regard to the information needed within the respective case. Saturation was determined by a) asking interviewees about referrals to highly relevant contacts within the wider organization and, following these, b) ensuring that at least one key person from data-related departments and one person linked to editorial was interviewed to end up with c) repeating patterns and diminishing returns from interviews towards the end of the data collection phase (Marshall, Cardon et al., 2013).[55]

[54] A meta-survey of the number of interviews conducted in qualitative dissertations shows a wide variation, with a median of 28 (Mason, 2010). However, lower numbers are also common (Wassermann, 2014). Other authors find saturation in nonprobabilistic sample sizes reached at around 12 interviews (Guest, Bunce & Johnson, pp. 74). In grounded theory, 20 to 30 individuals are recommended (Creswell & Poth, 2018, p. 226)

[55] During the later interviews at C4 and C6, I found repeating patterns in various ways, e.g. in terms of the types of metrics discussed, the different roles and job titles tasked with the organization of data work; the specifics of data infrastructures as well as the participants' general ideas around the subject of data.

Saturation was generally reached more quickly the more high-level contacts were involved.

Anonymity remains a critical ethical consideration in case study interviews (Saunders et al., 2015). In order to mitigate identification through detailed case descriptions, this study concentrates on cross-case synthesis. One way to support anonymity is to explicitly name all contacts inside of a comprehensive list, but to subsequently not attribute any one expression to a single contact (Yin, 2018, p. 297). Another solution would be to anonymize the participants alone while disclosing the entity under study (Yin, 2018, p. 298). However, this still carries a high risk of identification, especially if the number of interviews within the same organization is small or if different functionaries or hierarchical levels are interviewed. These considerations primarily concern the protection of the interview participants but also of the organizations providing sensitive information to the researcher as a whole. Overall, sensitive topics (such as strategic decisions, future plans, and internal disputes) could only be addressed here by first setting the stage of complete anonymity—both on a case and individual level. The assurance of complete anonymity was provided to increase the willingness of participants to share information about the organization.

Finally, when approaching potential interviewees, an interview guideline was provided ahead of time. This guideline serves as a trust-building measure with the interviewees. Before turning to the systematic development of the guideline, I want to reflect on the specific type of informant or expert interview that will be conducted for this study.

5.2.3 Semi-structured expert interviews

First we must clarify what is meant by the term "expert", a term historically associated with a wide variety of conceptions. Bogner and Menz describe experts as individuals who, through "their action orientations, knowledge and assessments decisively structure, or help to structure, conditions of actions of other actors" (Bogner & Menz, 2009, p. 54). In tracking the debate around the methodological foundations of the expert interview, they identify three distinct perspectives: a) the voluntaristic concept b) the constructivist variant of interpretation and c) the expert in terms of the sociology of knowledge (Bogner & Menz, 2009, pp. 48–53). In the present work, I follow the constructivist viewpoint on expertise, keeping in mind the sociology of knowledge perspective. I determine these experts externally, based on their ascribed status inside each organization and their professional (self-)positioning, assuming they carry expert knowledge of processes. As individuals tasked with generating and interpreting data, my experts fall rather precisely into this definition, as their whole work revolves around the ability to put into practice both their technical and interpretative knowledge (Bogner & Menz, 2009, p. 53).

In general, interviews are so deeply intervowen inside our "interview society" (Lindlof & Taylor, 2017, p. 170), it could be said that "the interview serves as a social technique for the public construction of the self" (Kvale & Brinkmann, 2014, p. 12). Expert interviews, a particular variation of the informant or elite interview (Lindlof & Taylor, 2017, p. 177–179), are among the most widely utilized qualitative research methods across various social science disciplines (Meuser & Nagel, 2009, p. 17).

As they are particulary used to capture and compare the knowledge of elites or experts from different institutions, such as corporations (Marshall & Rossman, 2016, pp. 174–175), these interviews are targeted at the expertise and knowledge of the informant-elite rather than their individual characteristics, conceptualizing them as representatives of a broader community (Johnson & Rowlands, 2012, p. 105). Particularly in an exploratory setting, expert interviews offer a means to gain focused and in-depth insights into the research area (Bogner & Menz, 2009, p. 46).

As an interviewer I will actively engage in expert discourse, acknowledging my background in the field at the onset of each interview while avoiding speaking as a peer. Especially considering interviews focused on work-related subject matter, an argumentative or discursive approach can be quite suitable (as opposed to largely refraining to intervene at all), as interviewees were found to expect the conversational structure they predominantly encounter in everyday situations from interviews (Trinczek, 2009, pp. 203–204). People in managerial positions in particular might project their everyday modes of interaction onto the interview situation; the researcher asks precise questions and they are the ones to answer precisely—just as subordinates or superiors would do (Trinczek, 2009, p. 204). Following this logic, passivity, non-intervention or rigorously structured questioning would potentially result more in irritation than in subjective free-flowing conversation with the interviewee. Additionally, sharing my interest in and understanding of specific expert knowledge might aid in building quick rapport (Lindlof & Taylor, 2017, pp. 193–194; Rubin & Rubin, 2011, p. 176; Bogner & Menz, 2009, p. 58).

Another objective of the expert-level approach is to standardize conditions across interviews as much as possible, mitigating any potential influence of personal acquaintanceship between the interviewer and the interviewees. The intention behind raising objections and presenting opposing views during an interview is not to persuade interviewees to alter their standpoints. Rather, it nudges them to comprehensively articulate and elaborate on their structures of relevance (Trinczek, 2010, pp. 211).[56] With that said, I still have to consider how the experts interviewed here do not all share the same "qualities", meaning their expert status was achieved under different workplace requirements and different levels of knowledge and their reactions or willingness to display candor also depends on these factors (Gläser & Laudel, 2010, p. 117).

In order to achieve a certain focus on the subject of inquiry, I partially structure my interviews through the use of a guideline or question stem. This semi-structured interviewing approach generally finds application when the researcher knows the general domain or topic of inquiry but does not anticipate specific answers (Morse, 2012, p. 199). The implementation of a guide serves multiple additional objectives: firstly, it helps to avoid appearing uninformed or unprepared in the presence of the interviewee, demonstrating a certain level of competence and attention to what is said (Kvale & Brinkmann, 2014, p. 134). Secondly, a guideline helps maintaining orientation and thematic focus throughout the interview process (Kvale & Brinkmann, 2014, pp. 133–134). Furthermore, researchers demonstrated how data gathered via a semi-structured approach becomes more reliable and consistent than data from completely unstructured interviews (Platt, 2012, p. 23). I discuss the conception and development of my interview guideline in Chapter 5.2.4.

[56] Towards the conclusion phase, depending on the level of candor established, I might confront interviewees by asking if they see data work as demonstrably contributing to operational goals (Gläser & Laudel, 2010, p. 149).

Expert interviews in institutional contexts are characterized by the fact that there are different perspectives on institutional action. At the same time, it is a group of people who, on the one hand, are practiced in expressing themselves and, on the other, see their own role in institutional action. For this reason, semi-structured interviews, or interviews with a uniform guide of questions are not entirely ideal for the task. While the guideline offers some structure and ensures a level of replication, I will prompt interviewees to provide consistent accounts of all relevant events from beginning to end in the sense of the narrative interview as conceptualized by Fritz Schütze (Schütze, 1983).

The interviews are structured in the phases of the introduction, narration, and inquiry phase and, finally, the interview conclusion. Part of the conclusion phase would involve the writing down of minutes by the interviewer. Immediately after the interview, such a protocol is completed and stored on a server together with the audio file of the interview and, if necessary, other files in order to prevent data loss. The narrative phase is about encouraging the interviewee to report using examples. Particular questions for working through these topics should be adapted to the interview situation in each case, and the order in which they are worked through may well vary. As this study heavily relies on interviewees providing in-depth descriptions of their activities, any usage of headlines or buzzwords will be challenged. If such language is used, I will follow-up and encourage the expert interviewees to give a more detailed account (Rubin & Rubin, 2011, p. 117–118). Whenever interviewees begin to narrate, I let them finish inside of their natural "Gestaltschließungszwang" (Rosenthal & Loch, 2002, pp. 221–222).[57]

[57] A compulsion to condense and detail repeatedly.

The demand phase centers on asking about points that are unclear in the above narratives or individual points in the conversation that were not addressed. The narrative phase and demand phase are not to be seen as sharply distinct from one another. As soon as the narrative part is finished, every narrative episode organically leads into an inquiry phase. At the end of each interview, I ask the interviewee if we failed to address important points in the area of the interview. This open-ended question often leads to another shorter or longer narrative episode.

Interviews take place as individual, face-to-face conversations and typically last one to one-and-a-half hours, so it is important to signal to the interviewee in the preliminary discussion that they should take sufficient time for the interviews. All interviews are digitally recorded via the conferencing application “Zoom”. It may be useful (if permitted) to take screenshots of visualizations or other material shared during the interview, e.g. websites. The goal here is to document events that evolve naturally out of the interview situation beyond the protocol to be filled out after the interview.

5.2.4 Interview guideline

The interview guideline for this study was designed to align with the theoretical building blocks established in Chapter 2 as well as reflect the semantic content of the research questions posed in Chapter 4.2. Each module of the interview guidelines corresponds to one or more theoretical constructs and research questions, aiming at a comprehensive exploration of the research topics.

From a structural point of view, the interview guide is composed of three main components: first, the introductory questions; second, the main section consisting of individual blocks of questions; and third, the concluding section with an outlook and a show of appreciation. With the objective of providing an introductory description of the phenomenon under investigation, the first block of questions centers around the interviewee's personal and professional positioning around data and data work. I follow general recommendations here and establish a common understanding around certain terminology (Podsakoff, MacKenzie, et al., 2003, p. 888). At the same time, these more personal questions aim to create a relaxed conversational atmosphere and give interviewees the opportunity to introduce their own ideas and perceptions.

The *first module* ("Personal background & role perception") primarily seeks to understand the interviewee's role and responsibilities within the respective media organization and get the conversational flow going in a casual manner. It allows for an examination of the structural and functional similarities and differences across different organizations in line with isomorphic qualities.

Module 2 ("Individual forms of data work") taps into the concepts of critical data studies (CDS) and *the data gaze.* Here, interviewees are asked about their use and interpretation of data and metrics in their daily routines, hopefully

providing insights into the power dynamics and ideologies embedded in data practices. On one hand, we delve into the social and political dimensions of data as recommended by CDS, on the other, questions about tools and visualizations involved in data work refer to the ways in which data shape the individuals' perceptive gaze and their actions.

With *module 3* ("Forms of data work in the organization") I further explore the potential of data and data affordances to transgress organizational boundaries. The questions aim to understand how data and metrics are used across different departments within the organization, highlighting isomorphic potentials. Additionally, the discussion of roles related to data and the changes in these roles over time also involves newly introduced work profiles and *boundary objects* which might be shared across the organization.

Linked to all four theoretical constructs, *module 4* ("Data work in general and look into the future") posits questions of historical and future developments in data work in media organizations. As I inquire into the influence of external companies on data work, I might shed light on the particular types of isomorphisms at play and the role of boundary objects in mediating these processes.

The *final module* ("Conclusion and reference procedure") does not directly correspond to any of the theoretical constructs, but it provides an opportunity for interviewees to reflect on the interview topics and suggest other potential interviewees. This can enhance the richness and diversity of the data collected, contributing to the robustness of the research findings.

To ensure comparability, interview questions were fleshed out as precisely as possible and worded in a neutral, open-ended and clear fashion (Gläser & Laudel, 2010, p. 135–144). As required, targeted follow-up questions were included to realign the conversation or delve deeper into specific details related to my main research interests.

In alignment with established practices, the interview guidelines have been progressively refined throughout the data collection process, particularly regarding the ordering and phrasing of questions (Yin, 2018, p. 62). In the context of data collection at the individual or case level, some circumstances may necessitate spontaneous adjustments to the process in order to gain a deeper understanding of the specific scenario (Eisenhardt, 1989). The efficacy of guideline-led interviews crucially depends on interpreting these guidelines as supportive tools rather than rigid frameworks (Gläser & Laudel, 2010, pp. 142–153). A significant challenge in conducting such interviews is the potential for interviewers to become overly rigid and excessively dependent on these guidelines, jeopardizing the natural flow of conversation (Gläser & Laudel, 2010, p. 146). To address this issue, I customized the interview guidelines to align with the specific role or internal organizational function of each interviewee in advance of the interview (Gläser & Laudel, 2010, p. 150). This flexible methodology permits the omission or modification of certain questions in response to the evolving discourse, thereby enabling the exploration of emerging themes or areas of particular interest to the interviewee (Gläser & Laudel, 2010).

In conclusion, the interview guidelines for this study have been designed to map onto the key theoretical building blocks of the research. Each module of the guidelines corresponds to one or more of these constructs, enabling a comprehensive exploration of the topics under investigation. By aligning the interview questions with the theoretical framework, relevant and meaningful data collection is ensured, enhancing the validity and reliability of the research findings. The initial interview guideline is provided as part of the appendix (Chapter 8.2).

5.3 Data collection and analysis

5.3.1 Study preparation and conduct

The gathering of qualitative data took place following the preliminary interviews spread over half a year from early December 2020 to mid-April 2021. The interviewees were formally requested in writing and were briefed by personal contacts of the author whenever possible. In individual cases, appointments were made through the personal assistants of the interviewees. In order to comply with the principle of informed consent, where participation in the study should be voluntary and its objectives well understood (Gläser & Laudel, 2010, p. 159), each personalized written request included a brief summary of study objectives and the methodology used, as well as the time horizon for data collection. Likewise, I gave the planned duration of the interview as one and a half hours, in line with common recommendations (Gläser & Laudel, 2010, pp. 162, 163).

An obvious necessity for expert or elite interviews, the interviewer needs to adapt to the interviewee's schedule (Yin, 2018, p. 85). Since the data collection phase took place during the COVID-19 pandemic, all interviews were conducted via videoconferencing using the virtual interview room feature of the software "Zoom". Similar to the case of classic telephone interviews, this resulted in advantages in terms of flexibility, time efficiency, and increased willingness of participants to engage in conversation (Gläser & Laudel, 2010, p. 153).[58]

[58] Notably, several individuals remarked on how the pandemic increased their willingness and availability to participate in the study.

While there are potential drawbacks in terms of the loss of nonverbal cues during virtual interviews, these are of secondary importance in the context of the current study relative to the information content derived from the conversation. Overall, the use of videoconferencing not only resulted in clear ecological advantages for destinations throughout Germany, but also made it easy to react to multiple postponements (Gläser & Laudel, 2010, p. 153). While meta-studies on interviewing techniques found remote interviews to potentially hinder a free-flowing mode of narration (Christmann, 2009, p. 157), these effects should be minimal here as I am almost exclusively dealing with professionals used to public speaking and video-conferencing.

In principle, it is a good idea to record the interviews for later comprehension and added accuracy (Yin, 2018, p. 109; Gläser & Laudel, 2010, p. 157). A full reliance on manual notes during the interview would not only irritate and interrupt the flow of the interviewee, but also risks loss of information or falsification (Brinkmann & Kvale, 2018, p. 108; Gläser & Laudel, 2010, pp. 157, 158). Before the interviews commenced, interviewees were asked for permission to record the interview as an audio track using the recording function of "Zoom"—again stating the purpose of the study and assuring anonymity (Gläser & Laudel, 2010, pp. 144). Only one study participant declined to be recorded, citing this as a condition set by their supervisor for consenting to the interview. In most cases, I otherwise felt that the formal setting of the interview did not create any increased reticence or bias. Only seldom was the sensitivity of the interview content reflected at all. At no point did the recording have to be interrupted, and there were no indications of confidentiality ("off-the-record") in what was said. This may be explained by the interviewee's familiarity with journalistic routines and their relative professional seniority.

Following general recommendations, interruptions of the interviewees were avoided at all costs (Brinkmann & Kvale, 2018, p. 83; Gläser & Laudel, 2010, p. 173). Important connecting points or follow-up questions were handwritten on the interview guide parallel to the statements and addressed after the statement was completed (Brinkmann & Kvale, 2018, p. 108). In order to maintain anonymity, no reference was made by name in the interviews to the other cases studied in each case and reference was also only made to interviewees within a case as an exception if they were known to the current interviewee as study participants anyway (Gläser & Laudel, 2010, p. 145). In the informal post-interview conversation, often new and interesting topics come up (Kvale & Brinkmann, 2014, p. 129). As this situation arose in one instance, I asked the interviewee for permission to include the topics we discussed.

5.3.2 Sample description

The cases examined for this study represent a wide array of different digital news businesses (see also Chapter 5.2.2), ranging from legacy media organizations to a fledgling startup. I present an overview of the cases and the respective case material in the following table (1). In the second column, format refers to the corresponding format(s) the organization has produced or still produces. Size refers to the business scale of the publishing house affiliated with the news brand, as defined by the OECD.[59]

Case	*Format*	*Size*
Case 1	Nationally distributed newspaper, digital	large
Case 2	Nationally distributed newspaper, digital	large
Case 3	Digital-only publication	small
Case 4	Regionally distributed newspaper, digital	large
Case 5	Nationally distributed magazine, digital	large
Case 6	Nationally distributed weekly, digital	large

Table 1: Characteristics of cases

Case 1 is a nationally distributed legacy newspaper, part of a larger publishing house in Germany. Case 2 is a nationally distributed newspaper brand inside a larger publishing house. Case 3 is a relatively new, digital-only news outlet produced by a small organization. Case 4 is a regional newspaper with limited distribution, that is embedded in a larger publishing house. Case 5 is a magazine with national distribution, that is embedded in a larger publishing house. Case 6 is a nationally distributed newspaper as part of a medium-sized publishing house.

[59] OECD, 2021, OECD SME and Entrepreneurship Outlook: 2021, OECD Paris, p. 142.

Overall, 21 one-on-one videoconferencing interviews were conducted, with only one exception to the modalities, as stated above. Out of these, one interview was a pre-interview testing the first draft of my guideline. The following table (2) illustrates the distribution of interviews as well as the career level of the interviewees.

Case	*Management*	*Intermediate*
Case 1	2	2
Case 2	3	0
Case 3	1	1
Case 4	2	1
Case 5	3	1
Case 6	3	1
Total	14	6

Table 2: Distribution of roles across the sample

To avoid identification, I substitute individual job titles where necessary, and use a broader list of roles when attributing speech to the interviewees inside the analysis. That means, whenever encountering unique or exceptional job titles such as a fictional "head of data greenhousing" or "data receptionist", I will utilize a more common equivalent based on the nature of their work, for example, "head of data" or "data scientist". Another measure taken was to obfuscate the genders of the interviewees by randomization.

5.3.3 Qualitative analysis

With the conclusion of all interviews, the audio material first needed to be transformed into text. These transcriptions were generated using a two-stage process: first, all audio files (in .mpa format) were fed into speech-to-text software provided by Happyscribe in batches.[60] While Happyscribe offers some form of speaker recognition and a certain level of sophistication, the resulting textual representations and labels (in .txt format) needed to be manually adjusted and corrected for punctuation, utterances, attribution, and orthography. Overall, the process took about an hour of work to produce the initial text files via automation. Subsequently, an additional one to two hours of manual work per recorded hour of audio was necessary to clean up the results. After transcription, the transcribed interview material was coded using the MaxQDA, a widely recognized software tool for qualitative data analysis (Kuckartz, 2019). The process involved structuring and coding the textual data into various themes, patterns and statistical relationships. This approach facilitated a comprehensive exploration of the material. As a guideline for my coding process, I adhered to the method of *thematic analysis* as reintroduced by Virginia Braun and Victoria Clarke in the field of psychology. This method is particularly notable for its accessibility and rigorous six-phase framework, which guides researchers from familiarization with the data through to the production of case reports (Braun & Clarke, 2006). This method also aligns closely with my intention to synthesize across cases and theoretical themes following the individual case reports. In compliance with thematic analysis guidelines, the following steps were reproduced (Braun & Clarke, 2006, pp. 92–104):

[60] https://www.happyscribe.com/

- *Phase 1: Transcription of verbal data*

  Rigorous and thorough orthographic transcript
- *Phase 2: Generating initial codes*

  Liberal and open coding of the material
- *Phase 3: Searching for themes*

  After the first iteration across all material, sort codes into themes
- *Phase 4: Reviewing themes*

  Check coherence of coded sets, reduce codes and themes to the essentials
- *Phase 5: Defining and naming themes*

  Describe themes, clarify their content and ensure clear separation
- *Phase 6: Producing the report*

  Produce reports on all themes with a clear analytical narrative

Following a distinction made by Braun and Clarke, I established latent themes (as opposed to semantic themes) underlying the material. This involved searching for insights into the reasons behind organizational change by asking about the "how", examining implicit characteristics of a professionalized data gaze, and identifying patterns of decision-making involving data. The development of the themes then involves interpretative work, and the analysis that is produced is "not just description, but is already theorized" (Braun & Clarke, 2006, p. 90). The coding process yielded a total of 1,597 coded segments, with an average of 93 codes per interview. Upon completion of all case reports and analyses, I revisited the interview segments used in the final body of text and translated these from German to English. But how can we reconcile this constructivist approach to coding with our more rigid case study design? On the surface, adopting a case study perspective informed by analytic generalization (Yin, 2018) seems at odds with a mode of analysis that leans on themes emerging from individual cases. In practice, though, Yin's method of

pattern matching, where the researcher compares predicted patterns (based on theoretical propositions) with observed patterns in the data, could very effectively be used to identify and interpret latent themes. Even closer to what Yin envisions, I will be generating latent themes that reflect "theoretically significant prepositions" (Yin, 2018, p. 141) as gathered at the outset of this study—resembling the analytical technique of "explanation building" (Yin, 2018). In addition, the often-cited rigidity in Yin's approach (as discussed in Chapter 5.2.1) primarily refers to the systematic nature of data collection and analysis, not a restriction on the depth or type of analysis.

While relying predominantly on expert interviews as my source of evidence (complimented by material provided by interviewees, corporate communications and job listings), through careful design of my interview guidelines and a precise line of questioning (See Chapter 5.2.4), a high level of construct validity was still upheld. However, it is important to acknowledge the limitations of the material in this regard. Furthermore, chains of evidence were established to maintain a clear and logical connection between the research questions, the collected data, and the conclusions drawn. Internal validity was ensured through pattern matching, a technique in which the observed empirical patterns are compared with predicted ones (Yin, 2018). This method allowed for a thorough examination of the data and the identification of any recurring themes or patterns. These patterns were then cross-referenced with the original hypotheses, allowing for a robust evaluation of the research questions. In terms of external validity, the study adhered to the principle of analytic generalization. This involved applying the findings from these particular case studies to broader theoretical propositions (Yin, 2018).

While case studies are often criticized for their lack of generalizability, Yin argues that the objective should be analytic generalization rather than statistical generalization (Yin, 2018, p. 58). Finally, reliability was ensured by creating a detailed case study protocol and maintaining a comprehensive database (inside of MaxQDA). The protocol served as a guide for the data collection process, ensuring consistency and minimizing the potential for bias. The database, on the other hand, provided a transparent and organized repository of all the collected data, making it accessible for future verification (Gibbert, Ruigrok & Wicki, 2008).

In conclusion, the analysis of the interview material was conducted in a rigorous and systematic manner, aligning with Yin's principles for case study research. First, making sense of the material and coding it along the six steps of the thematic analysis framework, using pattern matching to further condense the coding, constructing individual case reports and finally generalizing across my theoretical propositions. The use of the MaxQDA software for coding, combined with the application of Yin's guidelines for analysis, has allowed for a comprehensive and nuanced exploration of the interview material.

6. Empirical results

6.1 Overview

With this chapter, I now turn to the results of the empirical investigation. As discussed in the previous chapter, after individual case reports (Chapter 6.2), I then continue with cross-case findings (Chapter 6.3) along the theory-guided dimensions of inquiry in a generalizing manner and compare these findings with my original hypotheses.

Each case is illustrated with an organizational chart depicting data teams and data work functions (Eisenhardt & Graebner, 2007, p. 30). Starting with the organizational structure and an introduction to the actors and their role perceptions for each case, I then provide a relatively consistent set of subtopics that represent clusters of latent themes or specific threads of discussion that came up during the interviews.[61] My illustrative structure here could be described as mostly linear-analytic (Yin, 2018, p. 176). Subsequently, supported by empirical data material from each case, the results on the individual themes are analyzed in Chapter 6.3 (Eisenhardt & Graebner, 2007, p. 29). As I try to support each theme argumentatively with data material (Eisenhardt & Graebner, 2007, p. 29), I provide excerpts of passages from the interviewees. Since I am doing exploratory research, which usually aims to generate hypotheses for future research as well (Yin, 2018, p. 141), I also try to point towards new hypotheses throughout this chapter (Eisenhardt, 1989, p. 533).

[61] As an example, as privacy considerations were not explicitly asked about, a specific section on the topic only materialized for some cases. In other cases, demonstrations of dashboards by interviewees yielded enough material on the topic to warrant a separate section.

Some authors emphasize the importance of linking emerging theory from the cases with existing literature (Eisenhardt, 1989, p. 545). For example, Yin suggests following up individual case narratives with a chapter covering the cross-case analysis and results in a multiple-case study (Yin, 2018, p. 170). Since we regularly find results which deviate from the initially developed preliminary considerations in the course of a case study, thus requiring a revision (Benbasat, Goldstein & Mead, 1987, p. 373), in addition to a summary of findings across the theoretical dimension, I will synthesize results across all cases in the final chapter.

6.2 Case reports

6.2.1 The large publishing house (C1)

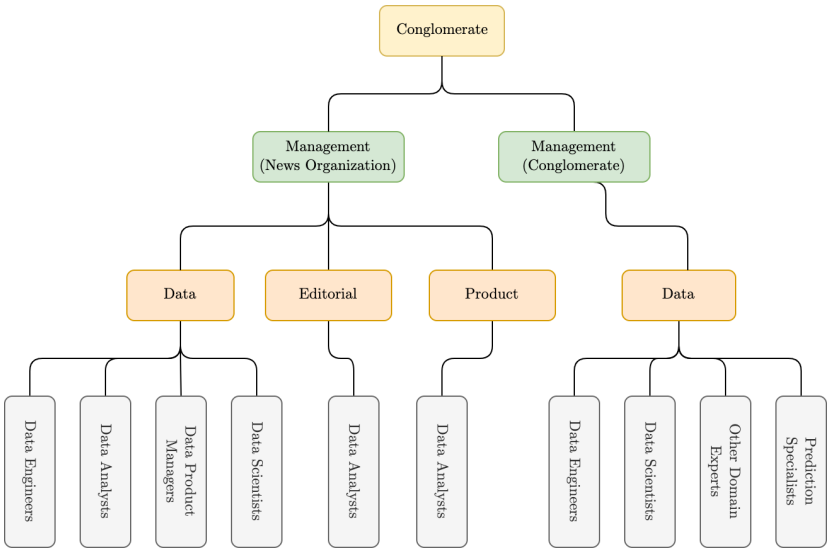

Fig. 1: Organizational chart of data work at C1

Organizational structure

In this case, three distinct areas of data work within the news organization can be identified, as well as one data team operating at the corporate level. In addition to an independent financial controlling unit as a common place of data work, the publication operates a dedicated data team tied to "product" and the operational pillar of "paid content", while other data work related to editorial analytics occurs separately and without its own organizational entity. There are no plans made for an integration of such roles and teams tasked with data work at the publication. Instead, efforts to centralize data work are made at the corporate level, aiming to turn repeatable data work into "services" available to all publications and editorial staff spread across multiple formats, brands, and publishers: "And now [there is] a third team [in addition to paid-

content and our data-team], which is trying to package our work into services across the group. It's not just [the newspaper brand], but also this entire group." (Interview C1-2, Pos. 222–225) Confirming this account, the data scientist sees the new data team as coordinating, "how data can best be utilized at group level or how certain data products can be built. Hence this division." (Interview C1-1, Pos. 223–226)

Metrics such as the *total number of subscribers*, a data and product manager recounts, have become measures of success for the entire enterprise: "There are targets on the business side which are then applied to the entire company and the primary goal would be the number of subscribers. We are not solely responsible for this goal." (Interview C1-2, Pos. 227–230) Notably, the data manager establishes a causal relationship between her work and the number of subscribers moving upwards. In doing so, she not only claims agency but regards her work to be on an equal footing with editorial work:

> "I think it is very much the content that plays a crucial role there. But we have the figures, make sure we report the right figures and also look at what is being done in the company and that might affect these figures." (Interview C1-2, Pos. 230–233)

As reflected by the data manager, the dual roles of measuring and reporting as well as optimizing for key metrics coalesce into a single responsibility. This creates a powerful agenda-setting device for upper management, while ensuring favorable results based on seemingly impartial data. Here, a relatively small data team of <5 data engineers and data scientists operates inside the publication—"data persons, as we like to call them" (Interview C1-2, Pos. 158–159). Interdepartmental demarcations are reflected in system access, with

CRM[62] still, in part, unavailable outside of marketing and sales. Analytics and tracking are primarily the concern of analysts, who are not part of the data team:

> "Unfortunately, we do not really have a CRM system either. Well, something along those lines with SAP[63], but the version number is very old. Many things are simply not possible. On top of that, my team does not have access to it at all, it's fenced off at the marketing and sales side." (Interview C1-2, Pos. 160–163)

Consequently, the scope of data work here remains limited to analysis, dashboarding, and some level of software development: "In our team, we wouldn't do any front-end development, but tasks in back-end development can also fall back on us. That means, for example, if we build a prediction model." (Interview C1-2, Pos. 318–322)

Still, the data team can technically cross reference sales and tracking data via the paywall software, which serves as a central data affordance available to all data personnel:

> "There is a synchronization with the SAP system because this paywall system now also needs to know about who has permission to read a certain paid article. [...] We extract a lot of data from there which has to do with inventory data. We also extract behavioral or movement data, meaning

[62] CRM is the "strategic process of selecting customers that a firm can most profitably serve and shaping interactions between a company and these customers. The ultimate goal is to optimize the current and future value of customers for the company." (Kumar & Reinartz, 2018); See also 8.1, "CRM"
[63] See also 8.1, "SAP"

which articles they see, from this T1 system, the CeleraOne[64] system, which also has the option of tracking, counting page views. We also look at that, but my team does not actually do that so much, it's more the analysts here." (Interview C1-2, Pos. 163–174)

These analysts, who are embedded in the editorial teams and first appeared to support editors with web analytics, have traditionally embodied a function of advocacy for editorial interests.[65] But there seems to be a shift in this relationship, with a reallocation of their work to the data team. The data manager hints at some discontent this migration of responsibilities caused:

"The analysts work closely with the editorial team and the business. And this project then came to us, from the analysts to my team, which is why there were a few difficulties last year as things that had been decided between the analysts and the editorial team had to be transported to us. And now we reached the point where we start digging deeper and involve the editorial teams more and ask them questions." (Interview C1-2, Pos. 288–295)

Despite the interaction with editors, sales and marketing remain the dominant influences on the data team and the key metrics under observation:

"Upper management, marketing and sales exert the biggest influence on us. [...] Management, because it is always a company goal to have X number of

[64] CeleraOne GmbH, founded in 2012 in Berlin, provided various software services to publishers, mostly recognized for their software related to "paywalls", was acquired by Axel Springer SE in 2019 to "expand technology and data competence" and "strengthen a strategic core area" (Axel Springer, 2019) and then sold to US-competitor Piano Software Inc. in 2021. See Chapter 8.1, "Celera One".

[65] See Chapter 3.5

> subscribers at the end of the year. There is naturally a lot of interest from the business side. In the meantime, the influence of digital subscribers and the overarching subscription business on the bottom lines for [publication] and [corporation] is increasing. Everything that has to do with paid content is a very strong focus and this attention is only growing." (Interview C1-2, Pos. 301–311)

Working together with the group's data team proves difficult, with monthly meetings often revolving around competency claims from the corporation:

> "[Meetings] are very difficult because there is also a lot of politics involved, rather at the levels higher up. That's why it's often a question of who does what and who is responsible for what. For example, they want to take more care of our content recommendations and we should take care of our subscription recommendations." (Interview C1-2, Pos. 351–355)

Technically, generating recommendations on a handful of pricing tiers appears comparatively trivial when stacked up against content recommendations. Splitting these recommendation concerns into multiple teams seems questionable from a steering perspective as well, the data manager objects: "I believe that these systems all belong together at this point. We are working on showing the subscriber or potential subscriber the right content and the right offer to bring them into a subscription and retain them there." (Interview C1-2, Pos. 370–373)

Overall, data and data work are subject to centralization efforts by non-editorial functions on multiple levels. On one hand, editorial analysis tasks are integrated into the data product team under the sales and marketing umbrella.

On the other hand, the establishment of two corporate-level data teams appears to diminish agency and responsibility at the publication level.

Actors and role perception

In this case, interviewees included not only dedicated data scientists and data managers but also an editorial analyst to represent the editorial perspective. With a background in computer science, the data manager at the news organization's data team (by contrast to a second data team at the corporate level) strongly identifies with her role as a technologist. She challenges the perceived exceptionalism of her field, imagining news production as a conveyor belt-type factory process—a mental model routinely applied to software development[66]:

> "Every industry feels that it represents something special and different from others. But you can actually learn a lot from other [industries] and draw parallels. [...] I believe that an editorial operation [...] is actually a Kanban process. It repeats itself again and again [...] I believe that a lot of what I have learned from previous software development can be applied here." (Interview C1-2, Pos. 420–432)

In the same team, the senior data scientist holds an advanced degree in natural sciences, where he first got in touch with advanced quantitative modeling. He also has prior working experience with a major ecommerce company.

[66] "Kanban" means "visual signal" or "card" and originated in the manufacturing industry in Japan to schedule or manage inventory levels and production processes. However, Kanban was later adopted by software development as an alternative to the traditional Waterfall methodology. In this context, Kanban signifies an approach to software development that emphasizes flexibility and so-called "continuous delivery". The rationale of this approach is to improve efficiency by only working on what is necessary at any given moment ("just in time") and delivering small increments of functionality frequently. See also 8.1, "Kanban".

As to their roles, the data scientist elaborates on how due to their small size and despite the staff nominally working with precise job titles, the data team effectively has to adapt and switch between contexts: "What I see here are these t-shaped skills. Everyone has their own areas of focus and specialties. But we try to position ourselves as broadly as possible. So that all areas of engineering, analytics and data science are covered." (Interview C1-1, Pos. 238–241)

A trained journalist not tied to the product management data team, the editorial analyst worked in several analytics roles in publishing before. He regards himself as an editorial-oriented equivalent to the analysts working in the product department, "where an analyst very often simply looks at data or conducts qualitative user research." (Interview C1-3, Pos. 7–11) Although embedded within the editorial department, the analyst remains on the payroll of the publishing company. Painting his personal motivation as helping editors make data-informed publishing decisions very clearly, he also acknowledges the economic goals of data work:

> "To help the editorial team, the individual editorial departments and individual authors to adapt and improve their own publishing with information on their readership. And to allow this information, which we gather through data about society, to flow into their own work as a data point alongside various others. The aim is always to have more readers, a more loyal readership and, of course, a paying readership." (Interview C1-3, Pos. 12–18)

A rare occurrence in the sample overall, the analyst here reflects on his own agency and power in doing data work multiple times. Aware of the inherent

interpretational margin of data, he defines himself more as a facilitator than a translator of data: "Personally, I try to steer clear of data-based interpretation in day-to-day business. Instead, I tend to talk to individual people, individual departments, individual teams in order to initiate or plan larger analyses, prepare analyses for these appointments and conduct debriefings." (Interview C1-3, Pos. 27–32) He also reflects on how his decisions in the creation of dashboards could potentially "nudge"[67] dashboard users towards decisions in a subtle or unintentional way:

> "These dashboards are assembled by me. I always start by making suggestions as I see fit. Of course, that's also a bit of nudging. [...] What a dashboard looks like or what information I can pull out of it naturally has a big influence on how a website is managed. [...] I can tell [editorial] in our all-hands meetings as often as I like how Google is an important traffic driver. That point won't come across as strongly as if they always have it in front of their eyes. And I can now decide whether I want to portray that." (Interview C1-3, Pos. 215–228)

Awareness of the inherent dangers of manipulative data affordances does not originate at the management level, the analyst states. Neither does management try to directly influence decision-making through the mechanism of data. Reflecting further on the issue, the analyst acknowledges dashboard building as his personal type of publishing power:

[67] Nudging is a concept developed by behavioral economists Richard Thaler and Cass Sunstein which understands designing choices as a way to encourage or "nudge" people towards making certain decisions. Nudging can be achieved by altering the context of choices, framing choices, providing feedback, or simplifying information to make decision-making easier (Thaler & Sunstein, 2008; Thaler, 2018).

> "I'm surprised that the editorial board or management doesn't use these tools to push editors in a certain direction. [...] Well, I definitely have an idea of journalism in the digital space and how to deal with data. And I present things in a dashboard in a way that I think makes sense. Yes, that's my back door for exerting influence without having to tell anyone." (Interview C1-3, Pos. 241–249)

On how he shaped his own role, the analyst recounts his early years at the publisher (5+ years ago) as a phase of establishing the data means to show evidence of the positive impact of his work:

> "In the end, I defined for myself that I need data and want to be provided with data so that I get a deeper understanding of my work. Otherwise I'm just optimizing out of the blue. And then we have endless debates about 'Alltagsevidenz' [68], which I simply didn't feel like having anymore." (Interview C1-3, Pos. 347–351)

Conversely, he considers the product department as the area "where it's about more money", explaining the relatively conservative role of editorial analysts in the greater scheme of data science and data work at his employer:

> "How professional is data work in editorial anyway? I personally can work with an Excel spreadsheet. When it comes to [statistical] significance, I can somehow wrap my head around that. But I can't do what a data scientist would traditionally do—actually see something and not just some stupid artifacts. [...] These people are situated in the product teams, where more money is at stake." (Interview C1-3, Pos. 562–568)

[68] Used here as a synonym for anecdotal evidence

Decision-making with data

We find multiple accounts of decision making based on data. The analyst shares his structural critique of current management and how it embraces data:

> "Management naturally believes data to be super important. Apparently they are enthralled by this data megatrend and the promise of salvation that data holds. [Management is] extremely keen to base as many decisions as possible on data. They do this because they often simply don't dare to make decisions based on conviction. And, when in doubt, they can't do much more than optimize KPIs." (Interview C1-3, Pos. 261–273)

Additionally, the analyst emphasizes how a shift in the profiles of journalism managers from industry specialists to general business or consultancy types prioritizes quantitative data informedness over more subjective decision-making:

> "In the past, management often consisted of people who came from a specific industry, such as the car guy or the publishing guy, who naturally also made decisions based on their experience. Now you often have people in management who have a classic business background or a consulting background. They come via the KPI track. They want to be provided with KPIs and optimize KPIs in case of doubt." (Interview C1-3, Pos. 261–273)

In terms of curatorial decision-making, the data manager suggests a hybrid approach, a collaborative model that utilizes technology to augment rather than replace human expertise:

> "I'm a fan of not letting algorithms make all the decisions. I'd rather ask the question of how best to combine people, editors and algorithms. [...] On our homepage, which is predominantly controlled by the editorial team, we have started a small section where we sometimes automatically decide the placement of articles, especially if you have several areas. [...] It's an area of three to four articles, and I think you can confidently have a mixture where you can say that part of it is editorial and part of it is decided algorithmically." (Interview C1-2, Pos. 28–39)

On the product level, data is used to inform and justify rather than to dictate which types of products to build next:

> "Then you have the product side and the digital products that we manufacture. Data also plays a very important role here because it can reduce complexity and help us make decisions, identify and then debate decisions in the first place, understand decisions and justify certain decisions. That's why data is of course also used very heavily." (Interview C1-3, Pos. 273–278)

Origins and changes

Ways of thinking about data and how metrics are relevant measures of economic success have changed significantly at the publisher. Changing sets of metrics introduced by the publisher over time reflect a progression through three distinct historical phases: Initially, advertising revenue was measured as a function of reach, representing the number of clicks or visitors on digital products. The publisher then experimented with a *composite metric* (a metric made up of multiple other metrics), before arriving at the current set of *key metrics* (field terminology to signify how a few select metrics represent the

operational goals more precisely than others): "In the past, people only looked at pageviews, regardless of where a pageview originated from. Then strange metrics appeared, such as the *time-on-page* in Chartbeat for the same situation." (Interview C1-3, Pos. 126–131)

While the composite metric still measured individual pieces of content—as opposed to subscribers, it reflected how content could increase reader loyalty:

> "To better reflect the journalistic model we aspired to, we then blended different metrics into one complex metric. Here we worked with [a technology service provider] and developed a content performance indicator, where reach, engagement and loyalty are incorporated as three factors in a composite metric. We moved away from this collective metric again. For a very simple reason. At best, the metric represented what corresponded with journalistic gut feeling anyway. However, it had the major disadvantage that it was difficult to break down and operationalize into individual measures." (Interview C1-3, Pos. 131–139)

All interviewees confirm that structural changes are underway at the publisher, with data work giving rise to entirely new units and roles in recent years:

> "We have a lot more traditional web analysts. We have people who are solely responsible for tracking, we have data scientist roles that didn't even exist three to five years ago. This job description may have existed elsewhere, but it didn't exist at all in the traditional publishing industry." (Interview C1-3, Pos. 311–314)

Older forms of statistical work and institutional knowledge on data from senior personnel are of little relevance to the new teams and roles that emerge today: "How did all this come about? I simply don't know. Because I haven't questioned much of our activities myself. But I think it's a shame that I don't know enough about it, because you could learn from what colleagues have done in the past." (Interview C1-2, Pos. 397–400)

But why have these changes occurred in the first place? Their emergence can be attributed, in part, to economic pressures. Skepticism towards data was overruled by a perceived need to measure business performance: "There used to be extreme skepticism towards data. [...] Now there are no longer these fundamental discussions about working with data, partly due to the economic pressure exerted on editors-in-chief which has trickled down to the editorial team." (Interview C1-3, Pos. 319–325)

The analyst also alludes to powerful data narratives as recounted by Silicon Valley corporations which publishers imitate:

> "As I mentioned, there is an ongoing data megatrend. A promise of salvation in data, that it could tell you something. Everyone knows these success stories from Silicon Valley, which were of course also pushed by the PR departments of said companies, saying how they achieved great success with data. That's why people work with data and like to reassure themselves with data and make this world more tangible, which they feel is always slipping through their fingers. And in the process you often forget to implement something concrete in your data work." (Interview C1-3, Pos. 286–291)

According to the data scientist, a fundamental difference between earlier phases of digital publishing and the present lies in the amount and quality of data available as "professionalism has grown proportionally to the amount of data we're working with." (Interview C1-1, Pos. 212–213) With the collection of proprietary raw data from an inherently data-producing medium like the internet, the publisher reached a tipping-point, where "data products" could be built:

> "It is perhaps an organic next step to work with more data, when you look at other sectors or industries. I remember when I started, we didn't have this interface on top of raw data either. Back then, you could do simple routine analysis but not really build data products. You needed raw data to do it." (Interview C1-1, Pos. 266–277)

On the contrary, even with more and more complex data, the analyst casts some doubt on the idea of data as more consequential compared to the predigital era:

> "[Data] is much cheaper and much more easily available to everyone. That's the big difference. Whether you draw a lot more conclusions from the data is a completely different question. I think that's the next step you have to take. To really work with the data. Not to fall into what they call a cargo cult[69]. To think that just because you look at data, you're doing well." (Interview C1–3, Pos. 436–442)

[69] Ethnographers suggested that "cargo" was often Western commercial goods and money, but it could also signify moral salvation, existential respect, or a proto-nationalistic, anti-colonial desire for political autonomy. See also 2.2.2

The notion of data as an increasingly important aspect of business seems pervasive, though not necessarily outside of management:

> "I have seen that in the last two years many more other departments are asking for figures or that figures are being sent around more frequently in reports. Mainly still at business level. And I would like to see this happening much more across teams. For them to know the current numbers. Not just their own, but also figures and numbers of other teams. Overall, though, a lot has happened." (Interview C1-2, Pos. 248–255)

In this sense, data workers increasingly advocate for the generation, access, and utilization of data as the ultimate objectives of these endeavors remain unclear. It appears as if data and data work have established a positive feedback loop: the more data is generated and measured, the more need for interpretation and utilization. As a consequence, sales teams and upper management exert greater control over these processes. The risks of running into a quantitative fallacy [70] do not escape the interviewees, wherein quantifiable objectives potentially overshadow other equally important goals, or even stifle innovation:

> "[Data work] naturally also has the consequence that other types of innovation have a difficult time in comparison. You quickly get into this mindset: If I want innovation or if I have to change something, then in case of doubt I have to be able to prove it with some KPI. If no KPI has changed, then nothing has changed. You can also measure yourself to death." (Interview C1-3, Pos. 455–465)

[70] See also Yankelovich, 1971.

With the concept of data science, a relatively new and controversial work profile was introduced between 2018 and 2020.[71] As the first hire in this capacity at the publisher, the data scientist recounts how awareness for the qualities of the emergent field was miniscule at first—the small team offered solutions in search of a problem:

> "Back then, we racked our brains and looked around to see what we could do with the data. The idea of a purchase propensity [machine learning] model was conceived collectively and over time it became increasingly important. There was a [2018–2020] project on product differentiation. [...] In this context, the model for calculating purchase propensity played an important role. And later, of course, management also sees that such a project is very useful or important in terms of our business model." (Interview C1-1, Pos. 85–103)

Casting some doubt on this assessment, the data manager notes how prediction and machine learning models in general are not yet in production, meaning they are not actively used: "At the moment, we don't have [machine learning models] in production that we offer as a service ourselves. So far, we have always integrated things into CeleraOne." (Interview C1-2, Pos. 318–324)

[71] The term data science originated in the 1970s, first as a proposal to rename "computer science", then as a substitute for "statistics" (Chipman & Joseph, 2016). In the 2010s, as a byproduct of general datafication and its pervasiveness at successful internet companies, the concept began to generate extreme attention ("hype") and today has reached wide adoption and is "more in demand than ever" (Davenport & Patil, 2022) across various industries. Although this study debates the inherent conflicts of the wording, from a field perspective, data science could be defined as "computer-assisted statistical methods on large amounts of digital data". See also 8.1, "Data science".

Data objectives

Looking forward, the data scientist identifies two major opportunities regarding data quality and centralization in firstly "better data quality and standardized data sources, the single-source-of-truth if you will, be it a data lake or a data warehouse" (Interview C1-1, Pos. 305–307). Additionally, the data repository should be "integrated with data from all possible sources. I think that's very important and we're not quite there yet" (Interview C1-1, Pos. 307–309). Here, the technically reasonable or even necessary thing to do also happens to align well with the organizational tendency, intentional or not, to place data power closer to the decision makers in upper management.

> "At the moment, there are still some silos for certain types of data, meaning newsletter or subscription data are very different things, also from a technical point of view. Such data gets stored in different systems which sometimes makes it difficult to define a common metric. It may be that different figures are reported from different sources for the same metric." (Interview C1-1, Pos. 314–323)

Arguably a problem sui generis, data fragmentation or heterogeneity is confronted with the powerful imaginary of a *data warehouse* (data lake)[72], a singular receptacle for all kinds of disparate data flowing through it, thought to obtain greater validity and harmonization in passing: "These problems will hopefully be solved by the data warehouse and we really only have a single source." (Interview C1-1, Pos. 314–323)

The editorial analyst's work has little to do with earlier iterations of this role, where editors themselves glanced over metrics like reach and abstract article

[72] See also 8.1, "Data Lake" and "Data Warehouse"

performance scores.[73] Now, the goal shifts to testing hypotheses over longer intervals and gaining a clearer understanding of the limitations of drawing conclusions solely from data:

> "My stated goal always remains to get the departments or the individual functional teams to notice certain things in their work with data on a daily, weekly or monthly basis and then simply develop theses together which we then collect. Team leads and department heads are responsible for collecting these theses and we then discuss them on a quarterly basis so that we have the opportunity to plan. What are we looking at anyway? Why are we looking at it? Can we adjust anything at all? Can't we adjust anything? And then carry out appropriate analyses." (Interview C1-3, Pos. 67–70)

A resulting analysis has to be long-running, because of the inherently multivariate nature of a digital journalism test environment, the analyst explains:

> "These analyses do not transpire like a classic product A/B test via a switch, where I send segments of users to version A and version B, but these analyses are [sequential]. Of course, and this is a problem with journalism, I am always dealing with many variables at a time. That's why I simply need to run tests across longer periods of time, because I can't always reproduce the same scenario." (Interview C1-3, Pos. 80–85)

In turn, the paradigm of live dashboards, an interface standard established by many years of working with editorial analytics, gives way to reports as a form

[73] See also 3.5

of delivery that matches the workflow. In this way, data work evolves into a more intermittent, interval-based practice:

> "This is obviously an automatism, I just have this dashboard and look at it. Then I have something like daily reports or weekly or monthly reports. And beyond these aspects, I think the most important work you have to do with data is to revisit it at regular and meaningful intervals. We defined this as a quarterly rhythm because there are simply an incredible number of variables that influence whether a text works or not." (Interview C1-3, Pos. 43–49)

Notably, the analyst here argues for data work to move away from constant monitoring and operating on data to a more distanced view on data—as the complexity of questions increase, it also takes longer to evaluate the significance or impact of changes.

Data literacy remains notably low at the publisher. Several interviewees advocate for education as a crucial avenue for fostering improvement in data work:

> "There are very, very few people at [publication], among the department heads, who I would describe as data literate. [...] How do you use data to sensibly adapt workflows, to adjust editorial evaluation a little? A lot of explanatory work [is needed]. I believe this is currently the most important task and will remain the most important task for quite some time." (Interview C1-3, Pos. 511–539)

Adding nuance to this, the data manager places responsibility on technical staff to educate and provide further elaboration: "What needs to happen more in data science and data engineering on the part of developers and scientists is to explain what you are doing. And secondly, to point out the possibilities. There remains a huge misunderstanding between departments." (Interview C1-1, Pos. 311–312)

In the analyst's mind, the concept of a singular conversion funnel[74] should be replaced with multiple ones that correspond to particularly successful "verticals". These verticals denote distinct information offerings that are specifically tailored to certain interest groups or topics: "I think the funnel exists too broadly. Addressing lots of users and then the rest gets forgotten. In some respects, this is also a bit of wasted energy. And I think we should work more in a vertical mindset." (Interview C1-3, Pos. 174–177) Editorial decisions regarding the creation of these so-called verticals could be enhanced with the use of data:

> "Whenever I see there is a lot of conversion from texts on the subject of psychology, but at the same time see I don't actually generate much reach, then I have to think about whether I can perhaps set up a kind of vertical service destination that I always keep fresh." (Interview C1-3, Pos. 168–172)

Data work should not become an end in itself, but always be tied to the concept of key metrics:

> "There's a difference between driving a car and knowing how to get from Munich to Hamburg as opposed to looking at the speedometer for eight

[74] See also 8.1, "Conversion Funnel"

hours. [...] We don't have to ask ourselves why something is moving [on the speedometer]. Nor do we need to investigate whether there is gravel on the road. These are not relevant questions. We have to ask ourselves how we can get from A to B with the lowest possible fuel consumption. These are completely different approaches." (Interview C1-3, Pos. 286–302)

Ongoing data work

In addition to the previously mentioned data reporting practices, which include ad hoc reports addressing specific inquiries from various stakeholders and regular reports sent automatically and repeatedly, the data team at the publisher can be described as being in an exploratory phase. As part of this exploration, the data manager envisions collecting more behavioral user data, aiming to understand the motivations of both current readers and potential readers. In her vision, the results obtained from these data points or measurements would align with editorial objectives as well—in this way, economic success metrics and editorial standards may converge: "We don't really know what's going on inside the reader or potential reader. [...] What parameters can be investigated? How can you actually express what the editorial team wants in numbers or something otherwise measurable?" (Interview C1-2, Pos. 55–59) Why would the team pursue the idea of more behavioral user data in the first place? Drawing on insights from the data manager's prior work experience in ecommerce, contributing factors for a "sale" would be measured much more broadly and all traces of a user would be constructed into a complex user profile to allow all kinds of deductions:

"I learned a lot about user behavior data, how to look at it and what kinds of metrics there are, especially in the area of engagement, which is still very much in its infancy for us. [...] We still do very little in this area. [...] There

> are all kinds of interactions. Not just the purchase of a product—you can ask a question, you can extend an offer there and give a rating here. We could measure every interaction that exists between the user and the product, and much more than we do now." (Interview C1-2, Pos. 406–441)

With the behavioral user data in hand, the publisher would then be able to identify distinct characteristics among several groups of customers:

> "One [project] is looking at how we measure long-term subscribers. It's called a cohort analysis, which is also the name of the project. We want to divide the users into different cohorts, i.e. the long-term users tend to stay and those who just take a quick glance at the subscription during the trial period [and leave]. How can we characterize these groups?" (Interview C1-2, Pos. 140–145)

As a cross-cutting requirement towards these goals, the data must first be discovered, then reach a level of completeness, and finally be transformed into a usable format: "We are investigating which possibilities there are for measuring and understanding the data and making it complete. That is one of the difficulties, to have data in the way we need it." (Interview C1-2, Pos. 152–155)

Data and the newsroom

Having worked in various data capacities at the publisher, the analyst recalls a previously negative sentiment towards data from editors, which has now somewhat diminished:

> "Questions were raised by traditional print journalists like: 'Why are we looking at it? Do we even need it? What does it tell me?' What's also interesting at this point is [...] that they don't want to be told anything by the data. Even if you were to explain that the data shows that this topic or that article simply doesn't work with the readership. That would be something which falls on deaf ears." (Interview C1-3, Pos. 329–334)

Confirming this perception, the data manager observes how negativity towards data could also stem from a loss of control over it: "I see many parallels in the skepticism towards digitalization and automation. The fact that you look very closely at metrics and figures, perhaps even handing over things that you actually want to control, especially in the editorial department." (Interview C1-2, Pos. 17–21)

Despite this air of negativity, since the introduction of editorial analytics as a data work concept over a decade ago[75], the editorial staff generally demonstrate the ability and willingness to work with even more data, dashboards and experimentation:

> "My first impression [<2020] was very positive from the get-go. Our stakeholders are essentially reader market and editorial team. I also had the feeling back then that they really enjoyed experimenting. [...] Before my time, our stakeholders had started working with dashboards and quantitative metrics, so it's not a foreign language and not new territory. That was my first impression." (Interview C1-1, Pos. 213–222)

[75] See also 3.5

As an example of current dashboard collaboration between the data and editorial teams, editors are encouraged to examine articles and their propensity to "convert", indicating the statistical likelihood of an article with a paywall directly leading a user to purchase a digital subscription:

> "Again, this [dashboard] is also about [paid content articles], how everything performs in terms of paid content. It's in real time, not quite real time, but updated every quarter of an hour. We show how often an article was clicked, where the users came from, [...] how many purchases or conversions resulted from this article. If you attribute that back, of course, it's not due to any one thing, we don't know for certain, but we consider that as the performance for the article, which then helps the editorial team place the article more prominently if it's doing quite well or to place it less prominently if it's not well received." (Interview C1-2, Pos. 271–282)

The main focus here is not on discerning the causality behind why any particular article ranks higher or lower on that scale. Instead, the emphasis lies on data-informed editorial curation aimed at driving increased sales. However, editorial autonomy over data is somewhat protected in this context. Work councils do not allow editors to be evaluated based on their article performance data, a practice that might be more accepted in other cases:

> "We are actually relatively free at the moment, but sometimes there are things that are more dictated by the business side: do this, work on that. Or when I hear: 'Please don't tell these figures to the editorial team, they don't want to know' or 'Measuring article performance is difficult for the works council' or something." (Interview C1-2, Pos. 496–501)

Aside from editorial analysts, there has not been an emergence of new roles within the editorial department with direct connections to data work, although data work has significance for various roles. This expanding editorial data work encompasses tasks such as traffic management, search engine optimization, and requesting and receiving data reports. (Interview C1-3, Pos. 359–360) On the technical front, editors lack the authority and expertise to make decisions regarding the acquisition or implementation of infrastructure or software:

> "In these matters, the publishing side or the people implementing things technically are very dominant. Of course they know certain tools, integrate them, stay in touch with certain service providers, hold the initial discussions and can talk to them on the same level. Traditional editors can't keep up. How could they?" (Interview C1-3, Pos. 400–404)

Indirectly, tools are picked and used in a way to fulfill the editors' expectations and standards:

> "When it comes to editorial data, I try to illustrate everything [with the same tool]. The reason being that we need a consistent shared vocabulary and the first big confusion always arises when two tools don't produce the exact same values. [Editors] are extremely hard-wired by their profession to find errors. And that's why you shouldn't ever offer something to them that could contain an error." (Interview C1-3, Pos. 102–109)

Overall, the editorial analyst's assessment suggests that the impact of data work on editorial operations at this publisher is relatively minimal. This may seem counterintuitive, considering his role at the intersection of data and editorial functions:

> "I feel that if the editorial team had no data at all, the product wouldn't be much worse. [...] It would be very similar to what we have today. I am strongly convinced here. In order to push content as an editorial product with the help of data, it would have to be worked with much more forcefully and consistently, in case of doubt also more reader-oriented." (Interview C1-3, Pos. 421–427)

Metrics and data sources

Here, editors are looking at *impressions by subscribers*, *impressions by non-subscribers*, and *conversion* metrics. In this context, impression simply means a single invocation or view of a page on any digital platform like a website or an app. Conversion is connected to the overarching notion of the conversion funnel, which originates from marketing.[76] In this framework, metrics are interconnected along an axis over time. The basic population flowing into the top of the so-called funnel gradually decreases as it "converts" through multiple stages, culminating in the desired outcome, such as a sale or customer acquisition, at the bottom of the funnel.[77] Such metrics inspired by ecommerce are increasingly relevant in editorial data work:

[76] The purchase funnel goes back to a marketing model first developed by advertising strategist Elmo Lewis in 1898 (Strong, 1925; p. 349f). Lewis charted the hypothetical route of a consumer from the point of being made aware of a brand or product to the moment of making an actual purchase. Later modified by both marketing consultants and scholars, the funnel serves as a mental model for guiding actions to optimize the various steps on said journey. Similarly, the conversion funnel as a term originated from ecommerce operations, where it refers to the route a buyer traverses while navigating a digital shopping platform before finalizing a purchase. See also 8.1 "Conversion Funnel".

[77] Conceptualizing a funnel can sometimes oversimplify or even distort the underlying system (or "product" in field terminology). In reducing the complex behaviors and interactions inside the measured system to a set of discrete steps with arbitrary metrics like "click-through rates" or "conversion rates", optimization towards these metrics might potentially overshadow other qualities of said system.

> "We have now reached the point where we picked out three metrics from this jumble or portfolio of available metrics. Metrics [...] which also represent the funnel, which has been used in ecommerce for a long time and is increasingly being used in editorial. [...] Views of subscribers are central to it all [...] and next to them are two other metrics. These are the page views of non-subscribers. Where can we reach new users? And that, of course, highlights potential on new [content channels to address potential subscribers]. And thirdly, conversions, where we can see by way of which individual texts we can get users to subscribe." (Interview C1-3, Pos. 139–152)

These metrics are believed to have a direct correlation with editorial output, suggesting that other metrics may not be associated with it: "These are three metrics that can be influenced in such a way that if you adjust something editorially, you should also see some effect." (Interview C1-3, Pos. 152–154) While the editorial analyst anticipates editors adjusting to these new metrics, the data team regards them as "standards" (Interview C1-2, Pos. 82) and demonstrates a more ambitious outlook. Their notion of the conversion funnel also extends beyond the subscription into retention metrics, measuring the longevity of relationships with paying subscribers:

> "We call it a value chain. Where do you run into a paywall at the beginning? [...] Then we have different subscription levels and then there's another retention level. It's not exactly defined, but we usually assume three months of paid subscription is the retention phase and after six months you're already a long-term subscriber to us." (Interview C1-2, Pos. 84–87)

The data scientist substantiates this focus on funnel metrics, and identifies subscriber *conversion rate*, *paywall click-through rate* and *retention rate* as the primary metrics that both the company and his team optimize for. (Interview C1-1, Pos. 108–125) After successfully getting users to subscribe, the most demanding and prevalent data problem seems to be the question of how to retain these customers: "After that it all gets a bit fuzzy, because that's where we really start to look deeper. These are exactly the things that happen after the closing [of a subscription]." (Interview C1-2, Pos. 117–134)

Overall, the data team works across four steps within the funnel, with their current focus directed towards retaining long-time subscribers: "Acquire users, retain users, and then acquire and retain subscribers. These are the four funnel stages. We are now working in these last two in the paid content area. We call the first two 'reach and engagement'." (Interview C1-2, Pos. 236–240) Regardless of specific metrics, all optimization efforts serve the purpose of increasing the overall number of paying subscribers: "The most crucial metric, I can say, that is interesting from the upper management side, is really the number of subscribers." (Interview C1-2, Pos. 100–102)

Technology and tools

Realtime-dashboards are predominantly used for what the analyst refers to as traffic management, similar to the operation of highway control systems:

> "Our dashboards appear to be relatively complicated and certainly overwhelm editors when they look at them for the first time. [...] It's important to me that they don't just look at these live dashboards because there's limited room for maneuver in a live situation. You can only really

do traffic management there. But you basically can't make medium or long-term adjustments." (Interview C1-3, Pos. 181–208)

The data scientist describes the production process leading up to the delivery of these dashboards:

"This whole dashboard project took quite some time and was quite laborious. We learned along the way that real-time data is not that easy to achieve. [...] There are also various options when it comes to visualization. The user first has to test or give feedback on how well a chart or visualization works. But what I have also learned is to apply the principle of 'eat your own dogfood', to be the user or play the user. Even before you go into user interviews with usability testing." (Interview C1-1, Pos. 171–189)

In technical terms, dashboards are based on *streaming data*[78], displayed through a managed[79] dashboarding service called Chartio:[80] "In [the real-time event handling software Apache Kafka] we can [run] real-time data. The data comes from Celera One [in the form of] log files on the server. The tracking events are then streamed into a Kafka cluster." (Interview C1-1, Pos. 192–203) Data aggregated in this way are then made available via the data sotware Big Query. [81]

[78] Acting on and passing of data as-you-go, as it appears at the source, as opposed to doing it at specific points in time in batches (e.g. once, at night). See also 8.1, "Streaming Data"
[79] "Managed service" refers to a model in which a third-party provider takes responsibility for some or all of a company's IT operations—in this case helping employees generate and share data dashboards. By outsourcing tasks to a managed service, a company typically reduces overhead costs from operating, securing and maintaining software or infrastructure.
[80] Chartio was acquired by the large, multi-national software company Atlassian. https://chartio.com/blog/atlassian/
[81] See also 8.1, "BigQuery"

As one of the few accounts of data work with machine learning (ML) in our sample, the data scientist speaks at length about a newly developed churn prediction model. "Churn", or "churn rate", describes the percentage of customers cancelling their subscription during a given period, such as a month or a quarter:[82]

> "It all started with an end-to-end[83] machine learning project. We started more or less from scratch. So we first created raw data, loaded raw data, tracking data, website usage data from our service providers. We then shoveled this [data] into the cloud, stored it temporarily and used it for feature engineering. This means that you can define and aggregate or convert certain user characteristics based on raw data. This would then be the input for the actual machine learning model. And the model uses these features and characteristics to try to predict what the user is likely to buy." (Interview C1-1, Pos. 192–203)

But, and this seems crucial to point out here, the resulting set of features to determine partitions of customers and their "propensity to buy", was integrated into what the scientist calls a rule-based system. This means that, rather than a ML model working in the background of a live environment, adapting and evolving, the experiment ultimately resulted in a piece of conventional software: "Ultimately, we extracted a rule, combinations of features, from this

[82] See also 8.1, "Churn"

[83] Meaning roughly "from start to finish", in the context of IT this could mean, for example, testing a software application with specific simulated user input and comparing the result against an expected outcome, the way the user would experience it. Another example would be end-to-end encryption, where message encryption happens on the sending device and only gets decrypted on the receiving user's device. Here, it means data generation, pre-processing, feature engineering, model training, prediction—all of which are performed automatically by the system. The goal of an end-to-end approach in ML is to minimize the need for human intervention.

model and implemented it in our paywall system. In other words, the whole thing is strongly rule-based." (Interview C1-1, Pos. 40–44) Across an underlying decision tree[84], customer properties appear like twigs on a tree, combine into branches representing desired outcomes, ultimately a new subscription. It remains unclear if, and to what extent, this model produces measurable results:

> "Each user is a data point in this multidimensional space with several features and you then try to partition the points in this space so that the data points or the target variable is as homogeneous as possible in a small partition. [...] The model is very interpretable, you can derive rules from it and then perhaps implement them separately elsewhere. There are disadvantages. Basically, you have to imagine that with the decision tree you start at the bottom, at the root. Depending on what features you have selected here first or how you partition it, this naturally has an influence on all subsequent decisions." (Interview C1-1, Pos. 63–74)

Among the tools used by the data team, CeleraOne and Google's Big Query are mentioned several times. Originally procured by the marketing team, CeleraOne allows to "segment" users: "Originally intended as a marketing tool, we now use it for data analysis as well. In other words, you can segment users with certain characteristics or properties into groups and then give them a name and see the size of the group" (Interview C1-2, Pos. 187–189) Big Query can be thought of as the central storage space for all kinds of data ("data warehouse"), which leads to some degree of lock-in into Google cloud products overall at the publisher. (Interview C1-2, Pos. 201–206)

[84] See also 8.1., „Decision Tree"; Fürnkranz, 2011.

In many instances, Google products are used to build internal dashboards for the data team as well, "often with Google Data Studio, with an interconnection to several other databases, which we can then create charts with free-of-charge." (Interview C1-2, Pos. 197–201) The data scientist seconds how the data team "essentially" uses Google Cloud, "otherwise, mainly open source tools and programming languages, Python and Scala. There are always different use cases, we use Python more for machine learning, to train models and for explorative analysis. [With Scala] we have also written tools ourselves, to download the tracking data from an interface for feature engineering or aggregating metrics." (Interview C1-1, Pos. 129–136) An analytics tool that seems to be primarily relevant to the newsroom, Linkpulse was not mentioned at all by the data team—a fact that could also be indicative of a shifting focus in data work from editorial analytics to custom reports, and dashboards owned by sales and marketing: "To get more granular [than with Google Analytics], we use Linkpulse. For example, to be able to say how clicks on a certain article came from a subscriber via the homepage." (Interview C1-3, Pos. 93–96)

6.2.2 The national daily (C2)

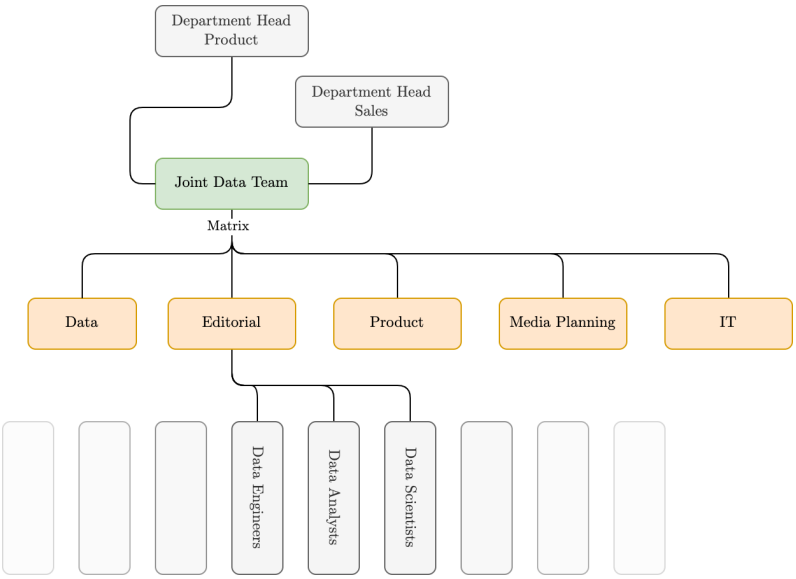

Fig. 2: Organizational chart of data work at C2

Organizational structure

As the only organization in the sample, in this instance, data analytics, data science and controlling functions are integrated into a single data department of <15 people. This department is overseen by the head of data who manages the joint data team (Fig. 2) and was established within the last <5 years by consolidating data work from multiple teams associated with digital sales (Interview C2-1, Pos. 39–43). According to the head of data, the responsibilities of the department include planning, setting up, and maintaining the data infrastructure; controlling; experimentation with data science methods; setup and maintenance of editorial or web analytics; providing reports and dashboards to other departments; test automation, as well as internal training programs and advocating for data literacy, methods, and quality control (Interview C2-1, Pos. 472).

As I will run into the data-adjacent concept of controlling in later cases as well, it makes sense to discuss its meaning here. While test automation, dashboards, reports, and analytics are covered elsewhere. Avoiding confusion, we need to differentiate between different concepts of controlling first. Controllership usually signifies the sum of all work carried out by the controllers in accounting and finance, which includes tasks such as "making cost information available, monitoring results and many other things."[85] In addition to this institutional, actor-based understanding of controllership (an example of which would be interviewee C5-4), controlling usually signifies a functional management perspective.[86] Earlier notions of controlling saw it as merely arbitraging information inside a business organization, limiting itself to an information-related dimension—with the goal of counteracting its "appearance of omnipotence" (Link, 1982).[87] A more up to date definition given by the economist Dietger Hahn views controlling as a management philosophy that "includes results-oriented planning and supervision by means of target agreements and analyses of goal achievement using the figures provided by accounting and finance".[88] In this sense, controlling very clearly represents a form of working with data and mirrors the cyclic notion of "build, measure, learn".[89]

Apart from the notable hierarchical aspect (controlling as a function of the data department), controlling does not entail accounting or finance in general here but instead focuses on subscriptions: "There is a separate controlling department. But they don't operate at a subscription level. They really

[85] Weber & Schäffer, 2008, pp. 2–3.
[86] Weber & Schäffer, 2008.
[87] Link, 1982, p. 261; Weber & Schäffer, 2008.
[88] Hahn, 1996.
[89] See also Bortolini et al., 2021.

specialize in revenue and cost. But when it comes to deriving meaningful values, it's not in controlling, but instead situated in sales." (Interview C2-1, Pos. 166–173)

All roles mentioned during the interviews have the functional equivalents of data analysts, software developers and data scientists, with only a single data scientist on the team. In addition to this, the publisher occasionally commissions external service partners, who "basically do AI as well". (Interview C2-1, Pos. 193–197) Similar to other cases, data science here carries the notion of a somewhat elevated form of data work—using statistical data and ML to predict certain outcomes:

> "We have one fully-fledged and one budding data scientist in our team who would like to go in this direction. And that's where topics such as churn prediction or the [paid] trial phase come in. To predict if and when a customer will convert or move from the trial phase to a paid phase. Clearly a matter of data science." (Interview C2-1, Pos. 97–103)

As was established, generating income through digital subscriptions receives highest priority at the organization—all interviewees directly or indirectly report to the sales executive. This emphasis corresponds with greater organizational complexity under the umbrella of the subscriptions division. Inside the division, individual teams work on three main concerns: a) the technical aspects of operating and adjusting the paywall[90], b) measuring the effectiveness of digital marketing actions, also known as *performance marketing*[91] and c) planning and measuring cross-media campaigns. (Interview

[90] A paywall is a digital barrier created by a news media organization that restricts access to their online content to paying subscribers. See also 8.1, "Paywall".
[91] See also 8.1, "Performance Marketing".

C2-3, Pos. 15–28) The main difference between b and c appears to lie in how performance marketing tasks are of a recurring or automated nature while campaigning deals with one-off tasks: "This team primarily concerns itself with offer-based campaigns, such as on topic XY with offer XY, while performance marketing is not quite so campaign-orientated, but [...] works in a continuous loop. SEA[92], remarketing and so on." (Interview C2-3, Pos. 27–28) Overall, the subscriptions "mission" (a terminology shared by other interviewees, e.g. Interview C2-3) has a higher headcount than others, "because there are substructures in the three teams. Above all, we take care of acquisition, recovery or the upselling and cross-selling" (Interview C2-3, Pos. 107–110).[93] Collaboration between the subscriptions and data departments happens on two levels, a) strategic and b) operational. In terms of strategy, the head of subscriptions imagines the dashboards provided by the data department as the device through which data helps facilitate control:

> "On the one hand, there is strategic collaboration, characterized above all by the fact that the [data department] provides us with dashboards. [...] We used to have Excel spreadsheets and now we have these nice dashboards. A great leverage for us and of course it makes our work much easier and also allows us to manage things much better." (Interview C2-3, Pos. 166–184)

Operationally, data analysts embedded in other teams gather and receive requirements that are not yet covered in dashboards:

---

[92] Short for "Search Engine Advertising", meaning the automated purchase of advertising space on particular results pages of search engines like Google or Microsoft Bing that match with the advertiser's goods or services.

[93] Upselling means to convince existing customers to move into a higher pricing tier. Cross-selling means to convince existing customers to buy additional products from the same vendor.

> "On a second level of cooperation we are very operational. I would not say that there is a directedness at all, but that it goes both ways. If the employees [of the data department] are embedded in other teams, there are of course [...] always new requirements. People constantly want to know things that are not yet mapped in dashboards." (Interview C2-3, Pos. 181–183)

Actors and role perception

In this case, all interviewees represent non-editorial functions and come from diverse professional and educational backgrounds. Coincidentally, their language contains a significant amount of technical jargon, which we need to clarify and deconstruct. Having previously worked in data science positions in other industries, the head of data identifies with the "revenue stream" generated exclusively from digital subscribers—making it clear that her team does not concern itself with advertising:

> "I started in sales management back then, that was the name of the department. By sales I always mean subscription sales. We had two revenue streams, advertising and subscriptions, and I was one hundred percent on the subscription side. [...] At the time, sales management included circulation reporting and all kinds of subscription analyses, and I also had my own analysts." (Interview C2-1, Pos. 13–25)

Not actively engaged in the discussion around data informedness, she prioritizes her mission to drive sales, "to make their decisions in a more data driven way" (Interview C2-1, Pos. 87–89). Even more drastically, her team aims to "provide data-based answers to all business questions—in the role of consultants." (Interview C2-1, Pos. 220–222). But what constitutes the

consultancy role here? Let us further deconstruct the narrations given by the analyst and understand the level of agency the data department possesses.

As an academic trained in the field of economics and marketing, the head of subscriptions reflects on how his work, the marketing of digital news products, could be categorized as a variation of so-called direct marketing: "Online marketing was already part of direct marketing back then [in my academic days], even if it has even more facets now than it did ten years ago. But at the core, we're doing the same thing. Now with the aim of gaining subscriptions." (Interview C2-3, Pos. 31–33) Notably, direct marketing could be regarded as a counterpart to advertising.[94] Where the former conceptualizes a targeted approach by organizations to communicate an offer to individual groups of customers, while also providing a mechanism for direct response (also known as direct response marketing), the latter takes a "scattershot", unidirectional, mass-message approach. In essence, the head of subscriptions here can be said to operate on data to facilitate direct or targeted marketing, which by definition generates data itself through response channels.

Before assuming his current role, the head of product held multiple technical positions in news organizations. Now, his team of [<20] product managers is deployed to other teams as part of "missions", overarching goals that product management deliberates and sets, to form the "connective tissue in the direction of the service providers building the apps and websites at the front-end".

[94] Lester Wunderman can be regarded as the "father of direct marketing"—his philosophy rested on the belief that consumers should receive personalized messages that resonate with them, instead of generic mass market ones. Wunderman's approach was inherently data-driven, involved an ongoing dialogue with customers and can be regarded as a precursor to what is now known as "targeted marketing" or "targeting" (Wunderman, 1994). Offline data and data mining were identified as requirements of direct marketing even before the wide-spread availability of the internet (Ling & Li, 1998).

(Interview C2-2, Pos. 15–19) He characterizes his role as one of mediation and knowledge transfer, where knowledge flows from sales through product management into the products themselves. (Interview C2-2, Pos. 18–19)

Origins and changes

Drawing on many years of technical experience in the field, the head of product sees the shift in focus from traffic to subscribers as a consequence of the volatility of traffic from search engines: "There's often a dependency, SEO traffic is a sweet poison. You take it, it tastes good and then Google changes something and the traffic has vanished again." (Interview C2-2, Pos. 306–309) In this sense, subscribers serve as a means of re-gaining autonomy and independence for the publisher: "The best way naturally is to retain subscribers. Direct traffic is the most valuable traffic of all." (Interview C2-2, Pos. 310–312) But how has the highly cost-intensive practice of search engine optimization (SEO) evolved to such significance in the field? The head of product thinks it is in part the result of news organizations mirroring their competition:

> "SEO became a huge topic in the 2000s and also in the 2010s. It went from being a secretive science for insiders to a broader discipline. Everyone was looking at one another and saying: they're using [SEO] to generate traffic, we need to do the same. That's how this run on traffic developed." (Interview C2-2, Pos. 334–340)

Apart from irritating behavior from search engines, other cultural changes paved the way for the shift towards subscribers:

> "This re-orientation to subscriptions also has a lot to do with technology. Back then, people did not yet trust payment methods. Ecommerce and

> Amazon were very much encouraging people to enter their credit card details, spend small amounts online or take out subscriptions. And only then did publishers have to find out what they could take money for." (Interview C2-2, Pos. 341–360)

In the early stages of transitioning towards a subscriber model, experimentation with paywalls required data and metrics that went beyond traffic: "Should we do metered models or hard paywalls?[95] Publishers first had to work their way to get answers, which is also very data-driven." (Interview C2-2, Pos. 341–360) In this sense, cultural changes led to technological changes before any form of directed organizational change management[96] could take place, a finding shared by the head of subscriptions:

> "We are right in the middle of a transformation. Data, metrics and KPIs are becoming increasingly important. [...] We are actually in the midst of technological change. This change also entails changes in data work. In our case, I would say that there has been a very strong cultural change at [publication] for the past [<5] years, particularly in the product and sales departments. And this cultural change has been followed by a technological change and finally an organizational change." (Interview C2-3, Pos. 277–286)

---

[95] A *metered* paywall is a subscription model for accessing digital news websites that offers a limited number of articles for free (in a moving monthly or weekly window). Once a user hits the limit, they are prompted to pay for a subscription. A hard paywall would require payment to access any content at all, whereas a *soft* paywall restricts access to select articles only. See also 8.1, "Paywall".

[96] Change management refers to strategies and techniques used by organizations to transition from their current state to a desired future state, as required in complex, dynamic corporate environments. These strategies acknowledge "the existence of employees as independent, acting beings" not for humanitarian reasons, but "aims above all at increasing economic efficiency." (Lauer, 2020; p. 5)

Reflecting the cultural shift towards product creation and sales within the organizational structure, a new department was established within the last <5 years. This restructuring saw the consolidation of longstanding marketing and sales divisions, which even retained historical ties to subscribers of print media in their nomenclature, into a single unified division. Similarly, all roles associated with data work were amalgamated into a dedicated data department. (Interview C2-1, Pos. 39–43) As the driving forces behind the decision to establish the data department, two executives from sales and digital marketing collaborated closely, highlighting the absence of editorial concerns during the conception phase: "A lot was initiated by the head of sales plus the former head of [digital advertising]. The two of them were actually instrumental in driving this forward." (Interview C2-1, Pos. 436–439)

What can be said about the specifics of this cultural change? There is a re-orientation towards the concept of customers and a professionalization in the degree to which these customers are surveyed and measured in more granularity around the concept of a "customer lifecycle". Instead of viewing a subscription as an "order that decides to take out a [subscription], there are always customers behind it. It's about processing these customers and moving away from the view that a sales department is solely responsible for producing converted subscriptions. Instead, we need to address the customer according to their phase in the customer life cycle." (Interview C2-1, Pos. 335–349)

Are these cultural (professionalization through customer and subscription-focused thinking) and technological (increased proprietary and granular data from reader research in digital media) changes truly novel? On the contrary,

the head of product sees it as a return to traditional news marketing using current technology:

> "Our price range used to be enormous, you had copy pricing, meaning single issue prices, then subscription pricing and also advertising. These were extremely good revenue models. Of course, the desire for subscriptions stems from the good old days. Because back then, people worked heavily with market research and sales data. [...] There has been a strong development towards a much more targeted approach. Subscription generation has adapted to the times and the possibilities." (Interview C2-2, Pos. 371–387)

On the other hand, these technological tools could also be viewed as significantly different from what was available before: "I think [the topic of data work] is very, very new. The possibilities we have now! [...] Everything is more up-to-date, faster-moving, more automated. In that sense it's not comparable with the past." (Interview C2-1, Pos. 481–484)

Taking a more neutral stance, the head of subscriptions reflects on the inherent novelty of allocating financial resources to data work: "The biggest difference is that in the pre-digital era, there were only ever the same marketing channels for several decades. Budget allocation was much easier to do than it is today. [...] Today, there is a much greater dynamic involved." (Interview C2-3, Pos. 349–355) At the same time, the publisher feels the pressure to compensate for losses in print subscriptions in the digital domain. As such, data work becomes a tool to facilitate the distribution of financial resources under pressure: "Publishers are definitely under pressure. For every print subscription lost, you have to gain three digital subscriptions." (Interview C2-1, Pos. 484–489)

Data objectives

There appear to be multiple competing objectives the publisher pursues with data. For example, as sketched out in the previous case report, data, and data work play an important role in experimenting with and selecting marketing options. This process helps inform decisions regarding the utilization of spaces within and around articles traditionally designated for advertising, aiming to maximize economic gains:

> "We have competing revenue models. For one, we have ad marketing, display advertising and then we have our subscriptions. [...] What's our goal here? Do we need [at the end of the article] a call to subscribe? In the same place, we could show advertising instead. So we have to constantly weigh our options and find the more lucrative option that contributes to our goals." (Interview C2-2, Pos. 45–53)

To reconcile these conflicting objectives, quantitative data are utilized for cost-benefit analysis, while normative data challenge the outcomes. Ultimately, this process leads to the reconfiguration of products:

> "We now have to cut up the page, the products, and consider at which point to hone in on which target, which KPI. And then coordinate. Of course there are editorial interests saying they don't want to have any marketing above our texts. In these situations, it's very important to work with data and track user interactions. What do they do? Do they end up in the funnel? Do they complete [a purchase]? Do they click? And is it all lucrative?" (Interview C2-2, Pos. 54–61)

In this sense, data carries argumentative weight in negotiations—in an automated fashion (the display space at the end of articles can be filled with varying items based on the individual user accessing the article) or in meetings (between marketing, product and editorial teams). But for data to have negotiating power, the negotiating parties would require some awareness of the availability of data and data affordances, what can be said or substantiated with it. Given that the objective of building data awareness was only established in recent years, it remains a critical challenge for data work to achieve success:

> "Tools get purchased, expensive tools that, worst case, are not used. We brought them in as a sort of fig-leaf. [...] But does awareness of what works and what doesn't in a data-driven way permeate all levels of the organization? It's no help if we create islands of specialists. [...] I would say that this has only been [resolved] for a few years." (Interview C2-2, Pos. 530–539)

Decision-making with data

Having entered the phase of organizational change, how do technical data affordances or data practices impact decision-making at the publisher? We find multiple variations of data informing decisions here: a) data-informedness as a cultural practice, b) data as an argumentative basis and c) data as a basis for algorithmic decisions. Within the context of data-informedness, all kinds of questions and initiatives are evaluated based on their impact on higher-level key performance indicators[97]:

[97] Key performance indicators (KPI) signify metrics that are used to measure and evaluate the success or efficiency of a project, organization, or individual against their set goals or objectives. Further discussion of the concept in case subsection "Metrics". See also 8.1, "Key Performance Indicators".

"Why this [digital] inventory space needs to be placed or why we need this revenue model or that feature; and why we might need editorial resources—there are no decisions without numbers." (Interview C2-2, Pos. 73–86) On the other hand, acknowledging their organization's reputation as an institution of quality journalism, multiple interviewees mention (unprompted) the importance of a clear separation between editorial and non-editorial functions ("firewall"). In theory, units like the data or product departments should not be able to make decisions that impact editorial work. However, data often serves as the argumentative basis for such decisions:

> "Experience is not enough. Simply claiming something works—you can't argue like that in a company like [publication], where work is very coordination-intensive. At [publication], we have an organization where editorial and publisher are separate in terms of content. That's what we call independent journalism." (Interview C2-2, Pos. 89–91)

Echoing this sentiment, the head of data asserts that data-informed decisions are poised to become the norm at the publisher, "so that exclusively data-informed decisions are being made" (Interview C2-1, Pos. 544–545). With data, the publisher would also be able to counter persistent habitual practices (implicitly attributed to the editorial personnel): "[You] get away from lines of reasoning such as: we've done it this way in the past, so we'll do it this way again." (Interview C2-1, Pos. 546–547)

Illustrating how data from controlled experiments might lead to incremental changes in the publisher's apps, the head of product recounts the introduction of a new audio player. Through the player, readers would be able to listen to a software-synthesized audio version of written articles:

> "We actually make data-based decisions for every product. Sometimes these are also me-too decisions or us-too decisions, meaning that if the market does it, we do it too. For example, integrating text-to-speech into articles. [...] The colleague who built and implemented this will say that there needs to be a huge play button above the text. And the editor says: let's put the button at the bottom instead. Then [A competitor] has a play button with the look and feel of Spotify and so forth." (Interview C2-2, Pos. 141–147)

To determine the best performing placement and presentation, each variant would be randomly displayed to users over a given timeframe and a specific contextual metric (the number of "plays" users initiated) would then be compared across variants—a system called "A/B" or "A/X" testing, depending on the number of variants.[98]

But how are these experiments rationalized? Personalization, the adaptation of digital offerings to individual users, is operationalized by observing time spent on said offerings—both audience goals and economic goals are seemingly in alignment:

> "The dream is to achieve this in a personalized way. How do we manage to offer the reader a custom environment, a periphery around the article that encourages them to stay, use additional features or click further? [...] For this we clearly need figures and A/B tests." (Interview C2-2, Pos. 163–165)

Results from these controlled tests regularly appear in presentations to executives:

[98] See also 8.1, "A/B Testing"

> "I receive reports with results from A/B tests prepared in presentations or decision templates. And these are then processed, put into context and I use them to make decisions or I take them with me and try to advocate for decisions at a higher level." (Interview C2-2, Pos. 178–183)

According to the head of product, such a practice of advocating for decisions with data reflects a hallmark of progressive leadership:

> "It is the job of managers like me, who come from this [technical] discipline, to signal they understand data and make transparent decisions. That wasn't always the case in the past because managers were basically still stuck in old ways of thinking." (Interview C2-2, Pos. 550–554)

As an example of algorithmic decisions, the publisher runs a ML model to calculate the probabilities of individual articles to persuade users into buying a subscription. If an article selected this way appears to be of general public interest, the algorithmic decision may be overruled by editors:

> "We also work with artificial intelligence and have a system that calculates probabilities and makes recommendations as to whether an article should be moved behind the paywall. And here it might happen that an individual editor or the editorial team decides to interject and not place a comment section behind the paywall." (Interview C2-3, Pos. 328–334)

Overall, there is some inconsistency in these statements, as data seems to be inherently tied to economic objectives or goals, in the form of key performance indicators, even if it concerns editorial work.

Ongoing data work

How are functions or responsibilities of data workers at the publisher evolving? In the year the interviews were conducted, the notion of "lead generation" as a success factor was introduced by the audience development division. Another concept from the field of (digital) marketing and economics, lead generation refers to the process of identifying and cultivating potential customers.[99] It inherently involves collection and action on *personally identifiable information* (PII)—to systematize these potential customers, also known as leads and allow for statistical inferences ("scoring") across multiple stages, again in the mental model of a funnel: "Lead generation is another topic that has moved up the agenda this year because it will be paramount to future growth." (Interview C2-1, Pos. 75–77) With data on potential customers fragmented across multiple systems, the data department's task is to provide a central database to systematize and evaluate leads. Notably, the head of data does not discuss strategies or plans to increase the number of leads:

> "To be honest, we are not yet well positioned on the system side. We are working hard on this. We have leads in different systems and don't manage them centrally; we don't have a centralized lead database. There's a target number of leads that should be generated in the entire [next year]. A lot depends on this metric." (Interview C2-1, Pos. 389–399)

Centralization efforts by the data department foster the concept of a "golden record" per customer, reminiscent of the idea of data as a *single-source-of-truth* in other contexts:

[99] See also Kumar & Reinartz, 2018. Systematizing leads in digital form usually happens inside of CRM systems, a concept discussed at various points in the present study. See also 8.1, "CRM" and "Leads"

> "In ten years' time, we want a 'golden record' of the customer, which we do not yet have. We currently still have different data pools, different data sources. We would prefer to bring them all together. Basically, I believe that behavior-based data will become even more important than it already is." (Interview C2-3, Pos. 384–391)

With the consolidation of behavioral user data (representing data gathered through user interactions on digital assets) and master data (representing legacy systems and static customer data), the interviewees expect to make entirely new assertions:

> "[One hot topic] is bringing together transactional data and base data in our data lake. This has been communicated as a vision in our team for [recent years]. Where we then bring together the usage data from [Adobe Target] with the base data [from legacy systems] and you can say whether certain things have an effect on shelf life, for example." (Interview C2-1, Pos. 111–127)

Proactively providing insights to dashboards and explaining their particular ways of measuring subscriptions with "clear planning and expectation cycles", the head of data goes beyond arbitraging information, instead his team appears to have an operational impact in identifying, naming and disseminating metrics. (Interview C2-1, Pos. 78–83)

Data and the newsroom

In line with the organization's status as one of the most trusted information sources in Germany[100], discussion about the potential tension between the

[100] See also 5.3.2, "Sample description".

data-informed subscription department and their editorial counterparts arose in all interviews. As the head of data puts it, "one has to handle editorial with a lot of tact and sensitivity". (Interview C2-1, Pos. 464)

One particular source of conflict arises in the overarching goal of subscriptions (some or all content obstructed by paywalls), as opposed to maximizing readership (all content publicly accessible):

> "Our goal as a [media brand] is to produce as many subscriptions as possible so that we can work economically. Of course, this is also a different objective to the one that perhaps prevailed in editorial in the past. In the past, the aim was to generate as much traffic as possible and a large readership. That's why it's not always easy for editors." (Interview C2-1, Pos. 458–463)

The significance of a certain data orthodoxy ("reine Datenlehre") might be even higher for competitors with a traffic-oriented business model. In the framing given here, a traffic-orientation approach leads to lower standards:

> "The higher your journalistic standards, the less impact pure data science has. [...] And there are certainly other companies that have less journalistic aspirations, that are more data-driven, that focus more on reach and clickbaiting. Data has an even greater impact on the question of what is actually published when, where and how than in a more journalistically driven organization." (Interview C2-3, Pos. 298–308)

Similarly, editors are neither evaluated by performance data nor do they work towards quotas established or tracked by the data department:

> "In the past, with reach-driven media that rely less on the paid approach, it was clearly about how many visits or page views or ad views an editor makes. [...] Many companies that are reach-driven now work in this way. Editors are given targets that have to be met or are considered a benchmark. However, this is not the standard in every publishing house. Others are more aggressive." (Interview C2-2, Pos. 61–67)

Editorial played "no central part" (Interview C2-1, Pos. 442–447) in planning and establishing the new data department. On the contrary, there was a general sense of adversarial sentiment. As the head of data recollects, the editorial team remained skeptical towards the undertaking: "They preferred things to stay the way they were. That's my feeling. Naturally, something has to be done and that's why they show understanding, but at the same time they were concerned about whether their interests are still being served." (Interview C2-1, Pos. 444–447)

In this case, the head of subscriptions actively steers discussion towards how data might challenge the editorial firewall:

> "It is absolutely essential that collaboration in today's digital world with the paywall in place is even tighter than in the past, when there was a product that was manufactured completely independently of sales and product development. And then sales took it and put it out there. [...] [In the past] there wasn't so much happening in the area of digital product development, whereas nowadays we are actually constantly developing the products with [our department], regardless of the content inside these products." (Interview C2-3, Pos. 267–275)

While acknowledging his influence and agency on editorial concerns, the interviewee also highlights how the conflict between editorial and publishing might arise from different success criteria:

> "We decide how content gets presented. That's why we need to work together more closely. Naturally, there is always potential for conflict, because our work [in our department] is more sales-oriented, while the editorial team works more journalistically and does not necessarily share the same success criteria." (Interview C2-3, Pos. 407–424)

External factors

There are multiple accounts of how external factors have shaped or inspired data work in its present form. Briefly, these are a) general technological advancements b) developments surrounding the subscription business model, c) the structuring and organization of data work, and d) the abundant availability of qualified candidates through a global supply-and-demand cycle.

The head of subscriptions explains how pioneering providers of digital content (e.g. Amazon, Spotify, Netflix) paved the way for digital news subscriptions by introducing consumers to online payment methods: "In principle, such companies increase the willingness to pay digitally." (Interview C2-3, Pos. 372–373) The executive goes on to assert that large digital platforms have influenced data work at the publisher on an organizational level rather than in terms of specific practices: "I wouldn't say there's a direct influence on data work here. But these companies have influenced our current organizational structure. And they also influence price points." (Interview C2-3, Pos. 377–380)

To explain the growing number of data workers in media companies, one should also factor in recent developments in the educational system, which have led to an influx of trained personnel. Combined with attractive salaries, this phenomenon fuels a growing supply-and-demand cycle:

> "Prices for SEO consulting have hardly changed, I would say. [The cost of] data has gone up. [...] This trend towards data scientists came iteratively afterwards. And that's where things are happening. Making a decision as a young professional, I often take the more lucrative route. I have a thing for data and can earn good money with my talent. There's a field of study, there are now degrees that didn't exist before." (Interview C2-2, Pos. 430–436)

With the professionalization of data workers, costs have increased dramatically: "Data work has become much more expensive. You have to consider SEO everywhere, everything has become more professional, you no longer work according to gut feeling, you no longer operate on the basis of assumptions." (Interview C2-2, Pos. 488–495) New, highly professionalized data workers introduce a new type of personality at the organization, an organized number-crunching attitude that was not particularly pronounced in previous generations of data workers:

> "[SEO experts] work with data out of necessity but are interested in the quick win. There used to be more of a 'Wild West' mentality: how can I achieve a great reputation and a lot of money with a few clever tricks? That's a very different approach than the number-driven person who deeply cares about the order of things." (Interview C2-2, Pos. 466–471)

Formerly, data workers often relied on guesswork and trial-and-error, especially dealing with search engines: "SEO tries to collect data that cannot be measured. [...] It's a bit like reading a crystal ball. SEO has to operate with a thin database because Google withholds search statistics. [...] You deduce a lot and make a lot of hypotheses." (Interview C2-2, Pos. 287–304)

Finally, it requires management decisions and receptiveness to admit these external impulses into the organization. In this case, first in the form of hiring consultants, then in hiring talent from outside the field: "Stimuli come very strongly from the market. You seek advice, you look at the market, you bring promising minds into the company. It's a multi-layered, complex process." (Interview C2-2, Pos. 564–566) Both upper management and the publisher herself are acknowledged to have been instrumental in this process: "The CDO has made a lot possible, and of course the managing director [...], who decided in favor of these people. [An editor] who thinks very digitally." (Interview C2-2, Pos. 578–585)

Metrics and data sources

As discussed earlier ("ongoing data work"), data at the organization falls into three broad categories, a) legacy customer data, b) external customer data and c) behavioral customer data. Such data exists in "various states and levels of aggregation" (Interview C2-3, Pos. 48), where aggregation refers to the summation of data points across arbitrary dimensions.[101] Data gathering on subscribers and marketing campaigns has been in existence even before the

[101] An example of aggregation of data points could be how "subscribers" would not be a useful metric in itself, but "subscribers across time", e.g. days or weeks, would make sense. The smallest available measurements across time are then "rolled-up" or "aggregated". Since the calculation of these time aggregated metrics can be computationally expensive, they are often automated in advance. See also 8.1, "Aggregation".

advent of the internet. Now, with digitalization, the overall data pool encompasses both legacy customer data and contemporary sources. External customer data would be data, which are "not in our own customer data, but are nevertheless accessible in our databases. For example, in the area of social media. There, we work on the basis of data [from advertising networks or third parties] and try to run suitable campaigns or acquire target groups." (Interview C2-2, Pos. 251–255) The third type would be behavioral data the publisher generates by tracking how users interact with digital affordances like the paywall. (Interview C2-3, Pos. 47)

From this data, several sets of metrics are constructed to serve as a foundation for observation, automation, and human decision-making. However, without integration into decision-making systems (whether autonomous or human), metrics are not inherently actionable. To prevent metrics from becoming ends in themselves, the key question then becomes how to operationalize them:

> "Customer lifetime value, for example, is a buzzword that everyone latches on to. But the question that needs to be asked, which few people do, is what to do with the number when you know it? How does it help me to know that there are two hundred Euros in a customer?" (Interview C2-1, Pos. 532–538)

Before adapting to the *customer lifecycle* as a mental model, the organization operated mostly on *key performance indicators* (KPI):

> "A lot has happened in the course of the organizational restructuring. When I started, it was mainly about inventory subscription data that lived inside of SAP and analyses on top of it. How many subscribers have we gained?

How many losses from which subscription categories? What does our quota and retention look like? In other words, these typical KPIs that you have in the subscription business, [...] but the task profile has expanded enormously since then." (Interview C2-1, Pos. 44-50)

In this case, customers are quantified across various waypoints (conversions) as they pass through the so-called conversion funnel.[102] But how do KPIs differ from other types of metrics? The head of subscriptions differentiates between metrics as general "principles and selection criteria according to which we work" (Interview C2-3, Pos. 89–90) while KPIs are those metrics intentionally selected as crucial to the businesses' survival and regularly reported to upper management. In particular, metrics considered as KPIs for the subscriptions team are *churn rate*[103], *cost-per-order*[104], *cost-per-interest*, *maximum-cost-per-order*[105] and the aforementioned conversions:

"Churn rate is a KPI just like CPO is a KPI. We have what is known as KPI reporting and this reporting contains certain, let's say, measurement quantities, such as costs for the CPI, the cost-per-interest or CPO, the cost-per-order or the responses to a specific advertising channel or a specific campaign. [...] And there is also a shadow variable in this KPI reporting, the so-called max CPO. This is a value that is defined as the maximum we

[102] See also 8.1, "Conversion Funnel"

[103] Concept also discussed in 6.3.1; See also 8.1, "Churn rate"

[104] *Cost-per-order* (CPO) as a metric signifies the average amount spent per order or sale as generated by a specific advertising action ("campaign"). It is calculated by dividing the total advertising spend by the number of orders generated by the campaign. By tracking the CPO, businesses aim to identify the effectiveness of different channels, messages, and audiences. See also 8.1, "Cost per order (CPO)".

[105] Although not a metric per se, Max-CPO establishes a maximum amount of Cost-Per-Order (CPO) which is deemed economical or acceptable. This limit is predetermined and functions as a benchmark. See also 8.1, "Cost per order (CPO)".

should achieve for the acquisition of a subscription." (Interview C2-3, Pos. 96–105)[106]

The head of subscriptions specifies that metrics are frequently employed within automation contexts, sometimes being both measured and acted upon in fully automated processes. An example would be harnessing behavioral metrics to determine individualized paywall offers to first-time visitors: "they might receive a different offer from us than someone who has already been there ten times and encounters the paywall for the tenth time. For me, that would be a metric that we use to actively manage." (Interview C2-3, Pos. 123–126)

On the executive level, two KPIs in particular are considered representative for success: contribution margin and the number of digital subscribers. With the former a function of the latter, executive management defines goals for digital subscribers as the overarching goal for the whole organization ("Gesamterfolg"):

"We have super specific targets that have been agreed upon and which we have to achieve. We should generate certain revenues from our subscriptions while not exceeding a certain budget. [...] A simple contribution margin, if you like. The second key figure is more strategically informed. Our goal at [company] is to achieve [<500,000] digital subscriptions by 2025. Naturally there's a somewhat more elaborated plan envisaging certain increases so that we reach this target. [...] That's the second key figure." (Interview C2-3, Pos. 140–154)

[106] Here, shadow quantity ("Schattengröße") refers to a metric that contextualizes another, often defining a desirable threshold. I assume that there are additional shadow quantities serving similar roles.

Departments began to self-regulate based on data and metrics as well, with the company-wide adoption of Objectives and Key Results (OKR), a management framework[107] recently gaining popularity among publishers.[108] The product department operates "with OKR on a quarterly basis. [...] There were already a few teams that did this before. Now we are doing it with the entire organization." (Interview C2-3, Pos. 189–194)

Technology and tools

How does the evolution of data tooling correspond with changes in data work, if at all? Here we find a shifting toolset, where the legacy systems are associated with certain undesirable characteristics. Older software gets described as rigid or inertial, often associated with giving arbitrary calculations:

> "When I started, we were very much focused on SAP MS/D[109], [...] a very rigid system, at least in publishing or at [publication]. That's how analyses of inventory data were made, mainly ad hoc work. And then data was pulled from the system via queries to [Microsoft] Excel and things were calculated in Excel. [...] The second task was reporting, which was updated in Excel at regular intervals and then visualized using [Microsoft] PowerPoint. That was actually a large part of the data work back then." (Interview C2-1, Pos. 410–412)

---

[107] Objectives and Key Results (OKR) is a goal-setting framework used by organizations. It comprises specific, measurable objectives and quantifiable metrics designed to track progress toward achieving these objectives. See also 8.1, "Objectives and Key Results (OKR)".

[108] For example, German news organization DER SPIEGEL introduced OKR in 2019 and described the process in a corporate blog: https://devspiegel.medium.com/okr-teams-kollaboration-wie-wir-unsere-produkte-weiterentwickeln-1190ac3fc055

[109] MS/D is short for "Media Sales and Distribution", a module or application component as part of the enterprise resource planning software (ERP) by German multinational SAP. https://help.sap.com/saphelp_nw73/helpdata/en/8e/c1865315b86359e10000000a174cb4/content.htm

In the ongoing discussion around software deployed in editorial analytics, it was revealed that the publisher used a free-of-charge version of Google Analytics in <2020. At that time, there was "hardly any expertise in analytics". Google Analytics was used to construct properties, "but neither in a structured way nor making any sense". (Interview C2-1, Pos. 418–421) The new data tooling is characterized by several qualities: Faster access to data, more up-to-date information, and the automation of data flowing from one system to the next. Overall, it is claimed how data workers can now make use of "daily updated metrics" to "simply construct dashboards". (Interview C2-1, Pos. 433) The new tools have "really enabled [the data department] to build a modern [business intelligence] infrastructure." (Interview C2-1, Pos. 418–433) Business intelligence (BI), encompasses the processes of analyzing and utilizing data to make informed business decisions. It often involves the use of dashboards, and interactive visualizations layered on top of data.[110] Expectations of modern BI include availability, accessibility, and ease-of-use for end-users. Echoing these ideals, the head of data describes how data was previously not easily accessible:

> "SAP cubes (storage units of data), for example, could only be queried on a weekly basis. This meant that, worst case, the data was a week old. [...] [You waited twenty minutes for the queries] and could go for a coffee. Now the data is retrieved daily and uploaded from SAP to Microsoft Azure, where it is automatically channeled through and used to feed dashboards." (Interview C2-1, Pos. 421–433)

Incurring significant expenses under the assumption of investing into data potential, the publisher went for top-of-the-line software options: "I think

[110] See also 8.1, "Business Intelligence".

[company] has reached for the top shelf here. And now it's time to utilize these possibilities." (Interview C2-1, Pos. 212–219)

With regard to editorial analytics, the publisher is currently in the process of discontinuing Google Analytics and consolidating all event tracking and editorial analytics into a single software, Adobe Analytics:

> "We purchased the big Adobe package and had the rollout [< 2020]. Before that, AT Internet was the tool that is now being switched off—and we still use Google Analytics. However, the aim is to switch to Adobe 100 percent. Because of course, there would be additional work involved in operating and testing two tracking tools. [We use] Adobe Analytics with [the AB testing software] Adobe Target. Adobe Launch is the tag manager[111] and Adobe Analytics is the system where the figures come in and you can build dashboards." (Interview C2-1, Pos. 63–70)

As described above, following the principles of business intelligence (BI), data should be accessible through dashboards[112], "data visualizations to support data-driven decision making" (Sarikaya et al., 2018, p. 682). In BI, they are "commonly more than a single-view reporting screen, a portal to the information needed for some goal and [they] may serve multiple analytical tasks." (Sarikaya et al., 2018, p. 685) In previous chapters, we established how dashboards are commonly distinguished by strategic, tactical and operational purposes. Here, dashboards outside of editorial analytics could be classified as

[111] Originally introduced by Google, now generally understood as a system to create and place snippets of logic ("tags") in the context of an app or website. Usually these "tags" are related to digital advertising, user behavior, or they *explicitly* or *implicitly* generate data. See also 8.1, "Tags / Tagging".
[112] See also 8.1, "Dashboards".

mostly operational in the sense that they are used to monitor systems on a daily basis, with ideal values always in view:

> "Firstly it's important that our dashboards differentiate between product and medium. [...] These are two important distinguishing criteria that for all dashboards, regardless of whether they are KPI, budget or circulation and revenue. The other distinguishing features depend heavily on the respective case. [...] As a rule, there is also still a distinction between planned and actual values. In general, there are always FAQs and glossaries." (Interview C2-3, Pos. 228–244)

Who uses dashboards across the different units at the organization? Top-level executives are less likely to incorporate dashboards into their routines, but "everyone on sales and product is looking at dashboards" (Interview C2-1, Pos. 309–311). The distinction between sales or product-related tasks and editorial purposes is also evident in the choice of tools, with Microsoft Power BI predominantly used for the former and Adobe Analytics for the latter. The focus on increasing complexity and investment in dashboards seem to be primarily directed towards the business side of operations: "Editorial dashboards contain other KPIs, web analytics KPIs such as page views, time-spent, conversions, top articles by conversions [...] Adobe Analytics is not quite as flexible as Power BI [...] The goal is not to further enrich Adobe [Analytics] with data." (Interview C2-1, Pos. 319–328) Dashboards are also associated with automation, as they offer self-service functionality, reducing the need for human interactions with the data department: "If the [data department] builds dashboards, trains all employees and they can answer their questions based on the dashboard, this will of course also save capacity in the medium term." (Interview C2-3, Pos. 206–222)

By contrast, application of a "deeper expertise" from the data department happens outside of these dashboards through "pro-active input." With more self-service dashboards in place, the head of product expects to have increased access to such expertise: "Although the [data department] is part of my team, as a cross-sectional department it doesn't work exclusively for us, but also for editorial. [...] That's why I would like to see even more proactive input. Because that's actually where the in-depth expertise lies." (Interview C2-3, Pos. 21–222) Confirming this perspective, the head of data sees dashboards as a means to automate repetitive manual data analysis and validation: "You don't need people who analyze reports somehow and tick off figures, because it all happens automatically on a daily basis." (Interview C2-1, Pos. 227–234)

6.2.3 The digital native startup (C3)

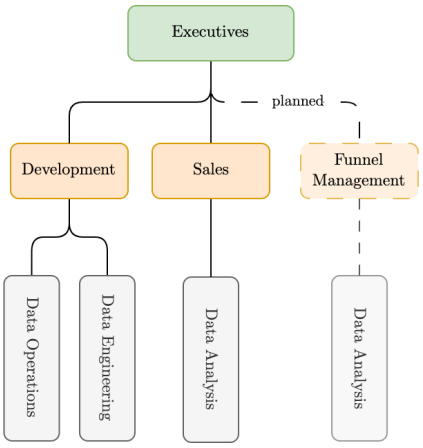

Fig. 3: Organizational chart of data work at C3

Organizational structure

At the digital native startup, with only <5 developers doing all the data infrastructure and data engineering work, the organization emphasizes journalism over technological innovation: "We couldn't afford any more. More developers would perhaps mean getting things done a little quicker, but it wouldn't immediately have a major impact on our sales, because journalism is already a very important part of our product." (Interview C3-1, Pos. 167–169) While there are no dedicated analytics or data functions at the organization, a salesperson handles these tasks together with the head of product and another executive (Fig. 3). In the future, part of these responsibilities will shift from the executives to a new role tasked with "optimizing the funnel and managing reach, newsletters, growth, and registrations in the long term. These are things that [we on the management board] have done more badly than well, but we are only starting to really invest money in them." (Interview C3–1, Pos. 503–505)

Actors and role perception

In the case of this comparatively small operation, data responsibilities are shared between multiple individuals who are primarily focused on other types of work, namely the head of product (C3-1) and the senior developer (C3-2). With a long-running background in journalism, the head of product also co-founded the organization. She speaks openly about the shortcomings her organization still faces in terms of technology and data, but thinks of improvisation as a virtue:

> "When I listen to myself, it all sounds incredibly professional. But it's not perfect at all! There's a lot of guesswork and 'Let's just take our word for it and go from there' and 'It won't be completely wrong' and so on. But that's okay. I have the feeling that we understand what's happening quite well." (Interview C3-1, Pos. 72–77)

As the builder and operator of all software at the company, the senior developer shares a similar mindset. In his view, the software stack might not be complex or sophisticated, but adequate and effective. In this sense, data work becomes subordinate to other tasks the senior developer has to fulfill: "Of all the day-to-day work I do at [medium], data work is just a small facet. [...] We are more like in the tiny data bracket." (Interview C3-2, Pos. 463)

Origins and changes

Initially, the startup adopted a membership-based model and embraced subscriptions without prioritizing data collection. However, they soon recognized the inadequacy of their data management practices: "we were shit in terms of data." (Interview C3-1, Pos. 235) Consequently, the organization sought to address this gap by acquiring the analytics software Mixpanel—

without truly comprehending its functionality or purpose: "[Mixpanel] wasn't a bad idea in principle, but we were really rubbish at using it. We didn't have the events in the right place, we didn't track [user] journeys properly and even then we wouldn't have known what it all meant." (Interview C3-1, Pos. 233–239) In digital marketing, a series of "events" form a data pattern used to make assumptions or deductions about individual users.[113] However, in this case, this pattern is broken: "We only understood the whole conversion metric after two years or so anyway, when we first understood our business model. There were no role models." (Interview C3-1, Pos. 239–241) Although the concept of conversions is consistently highlighted as a crucial metric throughout the sample, the head of product here pinpoints how it was still relatively new to the field in <2015.

After years of operation, the startup began to invest in data to better understand user motivations for cancelling subscriptions, "for about [<5 years]. That was the first time we collected large-scale user data from our members in order to playfully derive a few metrics to help us predict whether members were about to cancel their membership soon" (Interview C3-2, Pos. 47–51). Building upon previous analytics efforts, the startup began gathering data in larger quantities and with greater granularity (Interview C3-2, Pos. 87–93). This advancement brought with it new challenges in data harmonization and normalization, making it difficult to integrate data into routine practices or workflows:

[113] An event here refers to a specific action or interaction which is recorded or tracked on a website, app, or other digital platform. It could be triggered by a user interaction or behavior, such as clicking a button, making a purchase, watching a video or filling out a form. See also 8.1, "Event Analytics".

> "Naturally, you have the problem of being completely overwhelmed by this deluge of numbers that are totally unstructured and then you're convinced you have to make use of them somehow, but can't integrate them into meaningful workflows. It was an eternal struggle." (Interview C3-1, Pos. 94–95)

Once more, the startup encountered a lack of precedents to build upon in terms of workflows, which continue to remain volatile to this day: "There is no standard workflow, as with ecommerce platforms or something like that. I imagine they control all their metrics very well. We're not at that point here yet." (Interview C3-1, Pos. 733–739)

Having reached a local maximum of manageable data, the startup then decided to limit their scope to gathering data solely from authenticated users—meaning those users that have identifiably logged into the service:

> "We collected a lot of data for a while and reached our technical limits here. Not because we have too many readers. Rather because we collected too much data. So we went through several evolutionary stages to aggregate this data more intelligently and collect less redundant data. We've already invested a lot of time in tidying things up. Not even in real time, but rather downstream." (Interview C3-2, Pos. 121–127)

With this limited scope and the further aggregation[114] of data, the senior developer then achieved a manageable level of complexity: "We collect a lot of data and then run scripts to aggregate the data downstream. We usually have a daily granularity of actions." (Interview C3-2, Pos. 127–131)

[114] See also 8.1, "Aggregation".

Both interviewees emphasize how the nature of data work within their organization transcends the capabilities available before the internet era. Measurement fundamentally transforms as a consequence of the medium's interactivity:

> "Thinking of the GFK[115], that calculates ratings for broadcasters, you always need a feedback channel to be able to measure. [...] In this sense it's definitely something new, because there are completely different rules of the game. I can't imagine how you could have done something comparable in the pre-digital era." (Interview C3-2, Pos. 272–277)

On the other hand, while working with behavioral data and marketing funnels might be considered innovative in the field of journalism, these practices have travelled over from ecommerce:

> "I can't imagine that it was possible to operate like this in the pre-digital era because we have access to the behavior of individual users now. [...] On the other hand, we are not doing anything completely different than what's being done in ecommerce now. We optimize a marketing funnel and then there are sales, with costs in proportion to that [approach], CAC-to-LTV[116] or whatever they call it. It's no different for us." (Interview C3-1, Pos. 449–454)

[115] A German market research institution originally founded as Gesellschaft für Konsumforschung in 1934. Widely recognized for its commissioning to measure television ratings through custom devices called "telemeters".

[116] Acronyms for Customer Acquisition Cost and Lifetime Value, See also 8.1, "Customer Acquisition Cost (CAC)".

Rather disillusioned, the executive sees advanced data software and data scientists as a luxury reserved for the largest organizations in the field:

> "I don't see a future where many journalism companies can afford something like this. [...] There will, of course, be a few media outlets that are huge and form a sort of oligopoly like in the US. And together they bind all the big talent. It's almost like a winner-takes-all market. [...] This means that you simply can't afford a data scientist, because these are not tech companies shelling out money on expectations of high returns alone." (Interview C3-1, Pos. 449–452)

Editors might professionalize and cover tasks related to data work: "That's why I believe that journalism as a job has to develop to the point where you have to be able to do things like analyze data. It's not that incredibly complex either." (Interview C3-1, Pos. 452–454) Consequently, the head of product expects a consolidation of data service providers and even denies the existential argument of analytics companies like the widely used Piano[117]:

> "[In Piano] you can create these ultra-complex workflows in their backend. But it's all nonsense. It doesn't matter, we took a close look at it. Basically, Piano gives people an incredibly complex tool so that their steep prices are justified, but you could just as well operate without the complexity. In essence, it's a movement of concentration in terms of infrastructure towards a few service providers and a movement into the niche in terms of editorial content." (Interview C3-1, Pos. 591–594)

[117] Piano Software Inc. sells multiple software products to publishers around the topics of e.g. paywalls and analytics. Substantiating the interviewees' point about consolidation, Piano bought and merged with two other providers mentioned in the sample, AT Internet & CeleraOne.

Challenges

Facing constraints around ownership structure, business model, privacy, regulatory advances, and fundamental challenges with data governance, the startup faces several issues surrounding data. Due to its business model, paying members share a stake in the company as a cooperative, the startup needs to reconcile its data initiatives with its members' privacy interests. Targeting individual members for advertising purposes remains strictly prohibited by cooperative statutes and the team recognizes the operational risk involved with third-party services, who might not adhere to European privacy standards:

> "Naturally, we are challenged by the fact that our members don't want their personal data to be the product. I am more relaxed about it than our members. [...] Like any startup, we have to build a stack with a mix of dozens of SaaS [software-as-a-service] tools. [...] The costs are that you can't rule out the possibility of data being collected by external service providers. And secondly, that it will be passed on or even stored on servers that are not located in your jurisdiction." (Interview C3-1, Pos. 89–97)

Conversely, the step towards using database software from US-based company Airtable could be considered a milestone in its own right: "I thought the step of sending data to Airtable in the first place was a big step for [publication]. We regularly get feedback [...] about how important it is to members that their data is not used for advertising purposes." (Interview C3-2, Pos. 367–370)

Advances in data privacy also proved problematic in the way recent protections built into devices and browsers complicate the attribution of data to specific users:

> "Out of a hundred conversions, we can probably only attribute thirty of them to a clear source. For the other seventy, we don't know where the user was last. [...] We are trying to use heuristics to find the relevant page that we can assign to the subscriber as the reason for a subscription. And this deterioration in data quality due to increased privacy and data protection by browser manufacturers meant that we had to talk frequently about how we could improve the situation." (Interview C3-2, Pos. 242–251)

Initially, with the team unable to identify the root cause of the problem, editorial employees were left in the dark and editors had watched their "success metrics going down for months because these third-party cookie updates were only rolled out in batches [to browsers and devices]. We've noticed the metric drop a little every week and in that respect, our metrics are a bit broken." (Interview C3-1, Pos. 449–454)

Another contributing factor to broken metrics is the startup's culture of improvisation and experimentation, which included mishandling the database software, resulting in some costly trade-offs:

> "Airtable is essentially a spreadsheet based on a database [...] If someone renames a column in Airtable or changes a column type, the integration no longer works. As soon as a party changes something, [...] these changes have to be agreed on. That's the challenge with Airtable. Suddenly you spent half a day repairing things because something would always break. [...] In total, over the last two years, maintenance was expensive. If we had built the database ourselves, this wouldn't have happened. But then again our authors wouldn't have had the opportunity to experiment." (Interview C3-2, Pos. 318–333)

In essence, managing data and data work entails striking a balance between having enough data and avoiding an overwhelming torrent of information. It also involves navigating the challenges of operating with limited resources. As we have observed, the team struggled to navigate between the two extremes: on one hand, coping with overwhelming volumes of data that was technically unmanageable and exponentially more time consuming; on the other, contending with data insufficient to achieve statistical significance. For example, experimentation with pricing failed in this regard: "You need numbers to prove experiments and the hypotheses behind them. [...] In this regard we simply failed because we didn't have the quantity of signals or the expertise to be able to show statistical significance in our numbers." (Interview C3-2, Pos. 344–350) Furthermore, the team continues to deliberate over what to measure and how to correlate metrics with normative and economic goals:

> "We are caught in a paradox of choice. We no longer know what we should be looking at because we simply have so much stuff, both quantitative and qualitative metrics. But what should we actually want to pay attention to? What are the things that are really important to us and bring us closer to our goal?" (Interview C3-2, Pos. 281–288)

Metrics and data sources

As part of the membership model, the publisher has access to personal data voluntarily provided by its members. Essential tasks are performed based on this dataset, such as identifying experts on specialist topics among these members to collaborate with, or organizing physical events around the geographic locations of members:

> "[In the dataset] you can see that this woman works as a research assistant in religious studies at [a university]. So if [the editor] has a question about religion, she might be a source. And then she writes something that she knows a lot about, postcolonial perspectives, India [...] You can even look at a map to see where she lives. So if we travel somewhere, you can see who lives nearby and invite them for a beer." (Interview C3-1, Pos. 332–341)

The senior developer substantiates how crucial this membership data was for connecting with the community early on: "Airtable was super important initially, probably towards the end of [<2020]. To be able to fulfil the promise that we would connect with our members." (Interview C3-2, Pos. 156–157)

Out of all the metrics constructed from analytics data on digital assets, conversions are considered the most important: "Our most important metric, and we decided on it relatively early on, is conversion. That's a beautifully simple metric to begin with, because it's basically expressed in money." (Interview C3-1, Pos. 411–412) However, the head of product notes, the metric carries a certain risk of overinterpretation:

> "[With a conversion] it is only the last article that pushes users over the subscription hurdle [...] However, it is not at all clear whether this single article was really decisive for the new member. Instead, our analysis shows that it is a journey or relationship that develops over time." (Interview C3-1, Pos. 419–421)

Instead of focusing on individual articles, continuous "engagement"—routine visits and interactions by users over extended periods of time—might serve as a better indicator for conversion. But what exactly constitutes engagement in

this case? The team built a composite or coefficient metric across all the various traces users might leave:

> "We simply count when users open an article or set a bookmark or start an audio player or subscribe to a newsletter. All of these things, which have developed into features over the last few years, now have metrics that are tracked in the background. And, because the metrics were already there, we simply created coefficients from them." (Interview C3-2, Pos. 72–78)

With regard to achieving conversions through engagement, the startup draws conclusions about statistically probable paths that might or might not lead to a conversion. Users willing to take a survey, for example, have a high probability of becoming paying users at a later stage:

> "The more user engagement, the more truly active exchange there is between the user and us, the more likely the conversion is [...] The probability of someone filling out a survey and then becoming a paying member, thus bringing a lifetime value of 120 Euros or more, is around 10%. In other words, getting someone to complete a survey is worth a lot of cash. And that really impressed me when I saw the data." (Interview C3-1, Pos. 421–422)

Notably, the head of product does not use the engagement metric to illustrate the point about engagement but talks about the survey specifically. It appears fitting, then, how the senior engineer thinks of composite metrics like user engagement as inherently hard (and expensive) to construct. In his mind, said metric never passed the prototyping stage: "Especially this user engagement factor has been hacked together but it never developed beyond the prototype

stage. It's simply hard to justify the time and money." (Interview C3-2, Pos. 476–479)

An outlier in the sample, traffic and metrics around traffic numbers are irrelevant to the business model: "What many people say about us and what I like about our business model, is that we simply don't care about our traffic numbers. It doesn't matter how much traffic an article generates. As long as conversions are right, everything is fine." (Interview C3-1, Pos. 99–103)

Data objectives

Similar to previous cases, the primary objective pursued with data here is measuring users across funnels. We find a difference in nomenclature, with the stages labeled as flirts, followers, members and ambassadors: "Flirts are people who visit our site more than three days in a 30-day period. We recruit 90% of our new members from this pool or segment, while this segment only accounts for eight to ten per cent of traffic." (Interview C3-1, Pos. 216–218) "Followers" are returning users that have entered their email address to receive newsletters, where email addresses are necessary "to move away from this total dependency on large platforms" (Interview C3-1, Pos. 220–222). Newsletters play a crucial part in the conversion logic of the startup, with other metrics currently not observed as closely: "We see that it's much more important to measure this metric so then it becomes less of a problem if these other numbers are not as meaningful." (Interview C3-2, Pos. 264–265)

Data objectives are expected to shift in the future. As the growth in users will slow down or stall, metrics around the retention of existing customers will become more important:

> "We will eventually reach a point where growth is simply no longer proportional in terms of members, while the churn rate remains relatively constant and we reach a natural plateau. [...] It's my hope, that we will use data to make the product more interesting and more valuable when that happens." (Interview C3-2, Pos. 384–391)

At this point, the idea of a self-sustaining loop comes into play—with existing customers generating referrals through their personal network as so-called ambassadors: "We call them ambassadors, so that we can hopefully get a circular movement into the funnel at some point. But of course that's incredibly hard to achieve." (Interview C3-1, Pos. 223–225) Finally, with members sharing ownership in the startup, making financial data transparently available can be regarded as a core function of data work at this organization: "The idea is to make this transparent for everyone. Everyone in the company at least has the opportunity to get an idea of the financial development and, above all, the subscription development." (Interview C3-2, Pos. 203–206)

External factors

As stated multiple times by interviewees, they consider their organizational approach as pioneering in the field of journalism. Consequently, influence of other organizations or role models was deemed "miniscule" (Interview C3-2, Pos. 350). Rather than replicating big tech or conventional journalism platforms in terms of data work (considered a futile endeavor), inspiration is drawn from the field of ecommerce:

> "You have these highly perishable goods that you have to offer anew every day, which is why I think the comparison with ecommerce makes a lot of sense, because you can scale it somehow. Then you'll be Zalando at some

point, that's easier to achieve [then replicating technology platforms]. But the emergence of a globally relevant journalism player here in Germany will be a huge exception, if it happens at all." (Interview C3-1, Pos. 593–597)

Data and the newsroom

By foregoing a dedicated marketing staff and adopting a business model centered around the personal brands of individual journalists, the responsibility for data work in all its aspects falls upon the newsroom. This arrangement entails economic responsibility:

"The beauty of our business model is that the responsibility of our editorial staff cannot simply be shifted to a sales department or even an advertising department. In the past, it was necessary to put up an [organizational] firewall to prevent interference, but nowadays that is simply not necessary. Data work has no influence on journalism, except that it is more customer-orientated." (Interview C3-1, Pos. 581–583)

The membership database serves as a crucial research tool for all editors, enabling them, "to fulfil the promise that we get in touch with our members. If the editors want to do research and get feedback from the community, they should be able to do so in a proactive way" (Interview C3-2, Pos. 156–159). Other metrics data can be incorporated into the daily practices of editors and journalists working at the startup. Editors know their performance data; however management does not evaluate employees based on the number of memberships converted by their articles—which makes sense given the problematic attribution of conversions described above: "We can look at the memberships generated per editor, but we usually don't do that." (Interview C3-1, Pos. 512)

The head of product sees no data inclination, but instead a general frustration with data work from editors:

> "If I am relatively competent, data tells me what my customers, members and readers actually expect from me and whether they are satisfied. In everyday editorial work, however, this has not been realized in any form by us or other editorial departments. [...] At the big journalism players, writers are quite pessimistic and disillusioned about their future prospects—and rightly so. Dealing with data? I think they're happy if they even have a Twitter account." (Interview C3-1, Pos. 175–179)

Somewhat contradictory, editors should be the ones operating with data, not managers: "You don't need [managers] and they're just too costly. In my opinion, all this data work has to be done by the writers themselves somehow. So that the product has a solid foundation." (Interview C3-1, Pos. 183–185) Overall, these statements from the two interviewees about the impact of data practices on editorial work do not align well.

Technology and tools

Compared to other cases, the set of data technology at this organization is minimal. As the common data repository and interface used by all staff, data work centers around a singular piece of software, Airtable[118], a software-as-a-service database: "It's surprisingly good. Of course you have to clean up a bit from time to time, add new things, but then you can do quite a lot with the data that comes in. Right down to the level of individual users." (Interview C3-1, Pos. 178–179)

[118] See subsections "Challenges" and "Metrics and data sources".

All dashboards are built with Airtable as well, with both marketing and editorial functions sharing identical dashboards. Building proprietary data solutions would require specialists for operation, making development cost prohibitive:

> "Adding a tool like Kafka, Cassandra or Redis [...] to quickly cache or permanently persist large amounts of data would be a huge effort. Simply because you would have to acquire a lot of knowledge. [...] Our toolchain is very manageable and that is also very important, so that we could also put any given [developer] on it. We do very few exotic things." (Interview C3-2, Pos. 143–151)

One monolithic application[119] built with the Ruby on Rails[120] framework serves dual purposes: delivering the main product to end-users and interfacing with multiple SaaS data sources before consolidating data into Airtable: "Data sovereignty lies in the [Rails] application, which essentially delivers the front end and provides the external back end." (Interview C3-2, Pos. 108–110) Additional analytics systems like Matomo then become "just another channel", such as Mailchimp, which also provides newsletter metrics like opening rates. (Interview C3-2, Pos. 110–112)[121]

---

[119] Application monolith refers to a large, complex, and tightly integrated software application. It can be seen as a consolidation of various features and capabilities into a single, unified system. The opposite would be a software architecture consisting of multiple loosely-coupled services that are built and operated independently (also known as a microservice architecture).

[120] Ruby on Rails, often referred to as "Rails", is an open-source web application framework written in the Ruby programming language. Notably, it was first released in 2004 and developer interest has been declining for a number of consecutive years (Stack Overflow Developer Survey, 2023).

[121] See also 8.1, "Application Programming Interface (API)"

6.2.4 The regional publisher (C4)

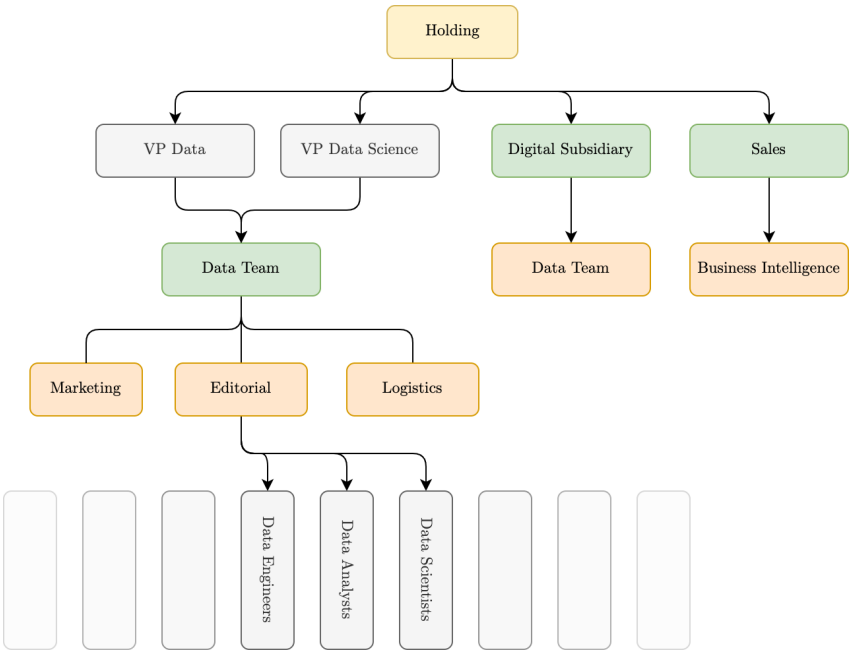

Fig. 4: Organizational chart of data work at C4

Organizational structure

At the regional publisher, management continues to experiment with different organizational figurations of data work to increase productivity, collaboration, and exchange across units. For this reason, the organizational chart presented in Fig. 4 should be regarded as a snapshot. Initially, data workers were embedded within other units in a matrix management model. However, the idea of possible overhead savings clashed with general confusion about these new roles and resulted in redundant efforts:

> "The transfer of know-how did not work as intended. And if someone only gets deployed to one area of responsibility of one division, they end up substituting and no longer take care of what's essential, namely data

analyses and empowerment of other people. Ultimately, we want to enable the entire company to work in a more data-driven way. That's why we decided in favor of an analyst hub [the data team in Fig. 4] in the current version. If we were to talk again in a year's time, I can say whether this was more successful." (Interview C4-2, Pos. 285–293)

As described by the executive, although the analyst hub still employs people dedicated to specific units, these people remain situated within the hub: "The analyst [would] spend one day a week with us, perhaps two or three days on-site elsewhere. Because 'what?' questions are identified very closely on-site in dialogue with the stakeholders. And the 'how?' questions are then dealt with by our team." (Interview C4-2, Pos. 304–308)

After the initial setup of a data lake, with the two senior staff in charge of dashboards and reports, a data engineer was introduced to the team.[122] Shortly thereafter, business analysts and data analysts followed (Interview C4-1, Pos. 200–209). Testament to a cultural change and to acknowledge its international team members, the data team now communicates exclusively in English: "In this team, at this local daily newspaper, which was unimaginable five or six years ago, the language we speak is English." (Interview C4-2, Pos. 509–511)

Similar to other cases investigated in this study, the publisher introduced data work as a relatively recent strategic objective with significant organizational implications. Following a period of passive digitalization, the executive explains, the company decided to redirect its focus towards audiences and data: "At some point we said that we no longer needed a traditional CDO. We have achieved digitalization. We need a new focus that is more orientated towards

[122] See also 8.1, "Data Lake" and "Data Warehouse"

audiences." (Interview C4-1, Pos. 66–69) The company then started a push towards digital subscriptions for "hyperlocal" and special interest content. This is also reflected in a new leadership team and the formation of a dedicated audience and metrics company: "We initiated this transformation [2–4 years ago], founded the [data company], brought new people into the company, changed the workflows in the editorial departments." (Interview C4-1, Pos. 59–62)

Who introduced the focus on audiences and hired the personnel to facilitate this change of perspective? We learn how initially, the publisher acknowledged the problem and decided to hire a new CEO with a background in empirical research and prior management knowledge in addressing fragmented audiences: "[The new CEO] carried out a study on what print readers and digital readers actually want to read. Overall the shareholder and publisher looked for people who he believed would make it happen." (Interview C4-1, Pos. 443–446) Additionally, the entire advisory board was replaced, with one board member advocating for the implementation of the Objectives and Key Results (OKR)[123] management framework: "With [board member] we are currently doing hardcore OKR development. We are trying to break down the corporate goal into OKRs, but these are interlinked for individual teams, some of which function very differently, meaning across [the technology, data and editorial departments]." (Interview C4-1, Pos. 459–464) The OKR tool as well as the performance data it displays, are provided by the data team, firmly establishing the team's alignment with the highest of management investments and goals.

[123] A management framework also discussed in other cases. See also 8.1, "Objectives and Key Results (OKR)".

Before the year 2020, the head of data explains, the data team was exclusively focused on delivering results that helped make the transition from an ads-driven business model towards paid content ("paid and premium"). This focus then gradually shifted to other responsibilities:

> "At the time, we were reinforced by IT to help us get to grips with the whole tracking world, because the focus in our area was clearly on supporting the digital transformation of our website to a paid-premium model. In the beginning, one hundred percent of our creative energy went into this topic and in the second year it was down to about fifty." (Interview C4-2, Pos. 69–74)

While the data team still mostly works with data after-the-fact, the head of data points to a more forward-looking role in the future: "[We don't just want to] somehow project key figures and historical values on walls or dashboards. That's what we're still doing right now. But ultimately, it's about climbing the data science or advanced analytics pyramid a little further." (Interview C4-2, Pos. 135–167) Yet at this stage, multiple interviewees view centralization and local authority over data as crucial to their efforts:

> "Our tasks include the implementation of new tracking tools, but also the integration of existing systems, reports or databases. [...] Because our philosophy or strategy in the data team is local data sovereignty, we strive for data sovereignty. We don't want to be users of a third-party provider or third-party tools [...] where you have flashing numbers and arrows pointing up and rockets and percentages—but you can't exactly say what it all means." (Interview C4-2, Pos. 160–167)

Here, third-party tools, while not entirely dismissed, are considered problematic in terms of explainability, transparency and user interfaces. Given the presence of data silos and functions across departments, the head of data and their teams should ideally report directly to executive management:

> "There are silos everywhere, colleagues collect and analyze data in every area, but only for the respective use case. I think it's important that we are an executive staff unit and not part of a digital unit or somehow subordinate to reader market [sales], but are able to act independently across all areas." (Interview C4-2, Pos. 76–85)

As many local publishers pursue digital subscriber growth, they increasingly measure and base their decisions on interoperable data and metrics. This trend makes their businesses more quantitatively comparable. As a result, a national initiative among local publishers has emerged to collectively pool their anonymized data and glean aggregate business insights:

> "I hope that [this initiative] will enable us to systematically test and learn together with other media companies that have a similar *North Star* and that we will make faster progress. By pooling our data anonymously and in compliance with the legal regulations of GDPR, we can make evaluations based on a completely different database than if we were to do it alone. [...] We formulate hypotheses together [in this initiative]. We are currently testing whether we are more or less successful with agency content, so-called evergreen content, as compared to others." (Interview C4-1, Pos. 555–568)

Here, the executive refers to the concept of a North Star metric, a quantitative measure that reflects the overall business goal of a company—often chosen to

align and motivate a workforce. As the executive points out, this initiative would have been considered unthinkable before. Under economic pressures, previously fierce competitors are now joining forces and share valuable insights:

> "The number of participants [in the initiative] to pool data in a large data lake and make it analyzable is now in the double digits. As someone who has been working in the industry for a long time, I'm just thrilled that something like this is possible. In the past, we were regarding each other purely as competition." (Interview C4-1, Pos. 573–577)

As the executive sees it, establishing a single overarching metric (the North Star) had a profoundly unifying effect on the company:

> "Individual editors work with it, hundreds of people. If the (data) system has unexpected downtime, we've now reached the point where people are calling all over the place and sound the alarm, because then they can't work. This is the kind of transformation we have undergone over the past year with [dashboarding and BI software] Tableau, the establishment and focus on above all a North Star metric and repeated explanatory work. It's actually quite simple: it's just this one metric, one so-called North Star, but we're all talking about it now." (Interview C4-2, Pos. 334–342)

Arguably, this metric mainly reflects a management strategy focused on measuring performance, rather than fostering a culture of data-informed decision-making or enhancing data literacy among the editorial team.

Actors and role perception

In the case of the regional publisher, interviews were conducted with all levels of management responsible for data-related matters. First, the publishing executive describes his role as both strategic and operational: "My focus lies on workflows, structures, strategies and expansion. [...] We need someone to organize [digitalization as a whole] and drive this change forward." (Interview C4-1, Pos. 34–56) Reporting directly to the executive, the head of data brings a background in operations from outside the journalism industry but has previously held various roles within the same publishing company. She describes her role as serving as an interface or translator between data science and stakeholders such as the editorial team:

> "I attend the editorial conference once a week, to explain the current state of affairs from a data perspective. [We] translate the whole thing into their language. They're not data analysts or data engineers, they're just editors or logisticians. That's a challenge. Incidentally, my main task is also to be the interface between those who use [data] and those who produce it. I can't actually do either. But ideally, I can mediate and create a common language." (Interview C4-2, Pos. 394–402)

Finally, the senior data scientist, with a background in engineering and prior experience across various jobs in resources and manufacturing, describes himself as scientifically minded: "I try to take a scientific approach to every single task in the company. [...] Based on my experience, I think it makes sense to do this. It allows us to see added value relatively quickly compared to simply analyzing an Excel spreadsheet and call it a day." (Interview C4-3, Pos. 23–28)

Data and the newsroom

The data scientist acts as an educator and gatekeeper for the editorial team, simplifying complexity and ensuring the comprehensibility of certain metrics:

> "I am glad we found this North Star. Because the metric is simple and hides away more complex correlations for the time being. In the background, we try to acknowledge and translate it all into a simple number. We realize that the editorial team is ready for more metrics. They are ready to understand more." (Interview C4-3, Pos. 263–268)

While this sounds rather patronizing, authority over performance objectives still lies with the editorial team: "Every time the targets are published, we look over them a few times, are involved, but instead of determining targets, we check their plausibility." (Interview C4-3, Pos. 334–336)

In straightforward terms, the executive asserts that the publisher prioritizes data-informed decisions over gut instinct: "I think that in the next few years, once we have learned to follow our gut feeling less and the data more, we will provide our colleagues with more data and develop our metrics more quickly." (Interview C4-1, Pos. 505–509) Contrary to the above statement, intuition is not completely overridden by data. Instead, the predominant principle in this ambivalent scenario remains a commitment to data-informedness:

> "I believe we have to be data informed and have an opinion as editors, act against the data or do something crazy from time to time. We have to move forward as a company, but of course we can't let ourselves be driven by the data alone." (Interview C4-1, Pos. 659–664)

The head of data substantiates the goal of data informedness, saying how the team would often hear "that the numbers don't add up and that instinct says something different" (Interview C4-2, Pos. 465–466) from editorial and that journalistic instinct will always remain a factor: "We're doing robot journalism now" (Interview C4-2, Pos. 466–469).

Metrics and data sources

Although *daily active subscribers*, the North Star metric, is not technically a composite or aggregate metric, it nonetheless entails significant dependencies. At the same time, the metric is easy to work with and understand—which to the executive is an important factor: "In the end, [the metric] means that you can communicate a clear focus to the editorial team and we don't get bogged down [by data]." (Interview C4-1, Pos. 305–309) The metric was developed in collaboration with a software consultancy and in close coordination with other regional publishers as part of the cross-organizational data initiative mentioned earlier. The overall goal of the initiative is to establish data interoperability between the publishers: "We are trying to make these publishers comparable. The other publishers don't have this metric yet. We are trying to link the daily active subscribers to everything that happens here." (Interview C4-3, Pos. 139–141)

The overall development of daily active subscribers is assessed weekly and takes center-field. Although not measured individually, per-team goals and performance are readily apparent to employees through weekly introspection sessions, where the reasons for missing or surpassing these objectives are discussed:

> "I look at this growth dashboard, a subscription dashboard, where I see how we are faring on a daily basis. We have our North Star report, which was published today. In this report, we publish every Monday how the individual teams are doing. [...] Today we have three teams that did not achieve their goals last week. More important than achieving targets, however, is what we learned from them. Every content team manager basically writes down every week what they take away from the week and what others can learn from it." (Interview C4-1, Pos. 524–537)

In the future, new metrics that work across different media types will be introduced at the publisher:

> "We will certainly continue to change our metrics over the next few years with the services we offer. One of the reasons why (the metric) *media time* is coming is because we are focusing more on moving images. We need interconnected metrics that work for video, audio and text. And we need our own metrics for video. We can learn from Netflix here. When are videos considered 'watched' and what makes a video successful? We have to define that for ourselves. We don't know yet." (Interview C4-1, Pos. 484–491)

More skeptical about the metric of media time, the data scientist argues for the exploration of more intricate metrics to qualify media time in the future: "Media time was introduced in the context of [data cooperation] as the one important metric. They only work with it [there]. I think it's a useful metric, but not necessarily the only one. You have to combine it with other numbers." (Interview C4-3, Pos. 291–293) In accurately modelling reader "pockets", metrics such as scrolling depth, read-through quota, and time spent on site, would be needed. Overall, the data scientist expects media time to better

represent the value a piece of content offers to customers: "In this way, we achieve even greater comparability by saying: great, [an article] has engaged one thousand readers for a total of one hour. That's the media time of this article. But a well-researched article that people read to the end is just as good." (Interview C4-3, Pos. 318–324)

During the initial phases of their data initiatives, the data team deliberated on and ultimately discarded a composite metric known as the *content performance score*, citing its limited explainability:

> "Some of the success [for a piece of content] comes from social media, some from Google, some from push notifications and so on. So you parameterize the entire event, weight it and you end up with a scale from zero to one hundred. [...] We later realized that an eighty percent rating is meaningless. What does that even mean for an editor? [...] We took advice from best practices and there were a lot of great approaches. We prepared everything behind the scenes to provide the calculations. But at the end of the day, we decided to only give the editorial team one figure as to not overwhelm them." (Interview C4-3, Pos. 269–285)

Reducing churn, defined as the percentage of cancelled subscriptions within a given timeframe, remains a critical operational objective in this case. While subscriber churn was already prevalent in the print era, the digital domain presents a more complex challenge: "Predictions about churn are a little different in the legacy area. Digital customers can leave every moment, whereas the intervals are significantly longer in the legacy area." (Interview C4-3, Pos. 457–459) The data scientist explains that prediction of churn works through a statistical method known as survival analysis. This method allows for the

identification of logical user segments based on their likelihood to "survive" over a period of time.[124] After experimenting with various combinations of time-series data points, humans can then label segments based on their characteristics. He elaborates on the contributing factors to customer attrition, particularly focusing on a segment of regular, highly engaged readers referred to as "brand lovers". Among his hypotheses is the idea that individuals who subscribe to newsletters are more likely to stay for a longer duration. (Interview C4-3, Pos. 416–424)

Origins and changes

Initially, the publisher embarked on their data initiatives with the aim of consolidating units and deriving insights from previously untapped data resources:

> "[At various points in the company] data was always talked about and then there was the awareness of data silos everywhere [...]. Our publisher was driven by the desire to know the name of the cats and dogs in households. We know the distance from the sidewalk to the letterbox to a centimeter. Why don't we do something with that information?" (Interview C4-2, Pos. 32–37)

In <2020, the team observed that the national publishing industry had been relatively inactive in its utilization of data. The emerging data team sought guidance from consultants, engaged with industry networks, and explored well established data strategies internationally, including those in Sweden, the Netherlands, and the USA: "Nobody was doing data [< 5] years ago. So we had to start somewhere and were very happy to find new communities next to the

[124] See also 8.1, "Survival Analysis"

BDZV. It all started with INMA." (Interview C4-2, Pos. 529–536) [125] At this early stage, the team formed under the label of data management, which was quickly dismissed:

> "The first official act on my part, together with [upper management], was to hire a data scientist. Who then said, [...] what you're planning is data science and analytics and then we quickly renamed our newly founded division data and analytics. That's much more accurate." (Interview C4-2, Pos. 37–51)

The insights gained through communities and networks continued to play a pivotal role for the publisher and its data efforts:

> "A strong momentum for us last year was [a digital transformation program]. [...] A bit of an action plan to work in an agile way and to achieve a clear north star focus. We had to defend and define our North Star. With the program, we took the step of looking more closely at audiences with a capital S, for example. We aim to have established a dozen audience teams by the end of the year." (Interview C4-1, Pos. 103–111)

Another core learning from the program was how to measure business performance in the digital world: "[The program] changed our work entirely. [...] In the past, we were successful when we had our standups in the newsroom every morning and complimented another on how good we were. [...] It was relatively easy in terms of metrics because you didn't have many." (Interview C4-1, Pos. 174–220)

[125] Acronym for the Federal Association of German Digital- and Newspaper Publishers, or Bundesverband Digitalpublisher und Zeitungsverleger. INMA is the acronym for International News Media Association.

In <2020, building on these lessons, management reframed the existing digital unit with the goal of aligning the company's digital sales and marketing under the North Star metric. While the unit still contains all the development resources and personnel, the name-change clearly places emphasis on its data work.

In addition, the publisher defined daily active subscribers as an operational metric, representing the number of users with a subscription who visit their digital offerings per day:

> "We can see that this metric means a major change of perspective in editorial. The phrase 'culture eats strategy for breakfast' comes up here from time to time. Which means we also have to manage change in this direction. We need to have colleagues who can survive this transformation in good health." (Interview C4-1, Pos. 192–195)

Overall, the executive seems unapologetic about this new performance culture, notably omitting considerations of how the competitive structure among teams might impact integrity, autonomy, and morale. "I believe we have come quite far in terms of mindset here at the company. We have implemented this focus on performance of our daily active subscribers very tightly." (Interview C4-1, Pos. 249–262)

Both efforts, a) identifying and catering to a growing number of smaller audiences and b) measuring the success of these audiences based on daily active subscribers feed into the overarching goal of reaching a certain total number of digital subscribers in the coming years.

Following the management decision to pursue digital subscriptions, the newly hired senior data scientist began exploring available data sources and prospecting data infrastructure with the aim of keeping data authority within the company (Interview C4-3, Pos. 16–18). He divides his narrative of data work into three phases spanning the last three years, with the first year dedicated to assessing and planning, scouting technology, and cleaning data:

> "We had lots of data from our production systems. The first task was to harmonize this data and make it accessible. One person receives one report, another receives a different one. They might be looking at the same thing, but have two different numbers. Why? That's your typical story. So we decided to take the data lake approach. [...] So that we end up with data models that are standardized and cover many areas, a so-called one source of truth." (Interview C4-3, Pos. 44–61)

In their first year, the data scientist and the head of data divided their work between technology and gathering requirements (Interview C4-3, Pos. 66–69). Building the data lake was a time-consuming task that ultimately took an entire year (Interview C4-2, Pos. 76–85). Purchasing external technology services ultimately proved unsuitable in the initial phase of the undertaking:

> "There are lots of startups eager to do this work for you. But I've been dealing with the problem of cleaning up data for three years now. I can't imagine a service provider could do it faster than us. Maybe better, but not faster. They don't get our business logic." (Interview C4-3, Pos. 71–92).

One test run with a data startup was eventually cancelled because the resulting data "somehow did not look right" and the team wanted to move away from being "married to service providers" (Interview C4-3, Pos. 71–92).

The second year was marked by the expansion of the data infrastructure and delivering results under mounting financial pressure:

> "In the first year, management realized that we first had to structure and plan. From the second year onwards, they approached us and wanted to know what was happening. There was a lot of interest from editorial and the digital marketing department. [...] They wanted to see performance figures on editorial [...] This has not been trivial." (Interview C4-3, Pos. 105–115)

The data team then began to deliver editorial dashboards showcasing analytics data created using Tableau. Here, the head of data draws an analogy to logistics:

> "If you make everything available to everyone [...] then all editors turn into analysts. That was a noble idea, but about as realistic as Facebook winning the Nobel Peace Prize for data protection. It didn't work. That's why we said we needed a tool that provides data for the respective user at the right time, which brings us to logistics. The right time, the right quantity, the right quality, the right data." (Interview C4-2, Pos. 211–219)

Early experimentation with metrics showed mixed results, as article-based performance scores hardly correlated with digital subscriptions: "Here the first assumption was to look at the number of registrations for an article. We quickly

realized that not all articles lead to a registration and the articles that do, surprisingly are no better than others." (Interview C4-3, Pos. 121–128)

During the interviews, the ongoing third year was characterized by experimentation with machine learning (ML), and research on the fundamental concepts of and prediction[126] and natural language processing (NLP):

> "We said, we're ready, we can do more than just map the status quo. We can implement simple approaches to recognize entities from article texts for example. Such tagging is a well-known problem across all media companies. Sometimes tagging is very structured and clean, sometimes it's an afterthought and in some cases it's not taken care of at all. We have tried to find an approach where it does not necessarily have to be part of the editors' workflow. Instead we want to automate it." (Interview C4-3, Pos. 158–165)

How has digital data work evolved compared to the marketing of print subscriptions and products? "In earlier days, colleagues in reader market already worked with personas and had a clear idea of [the core target group]. But the core business in reader market, was to call and acquire customers who were mostly known. So it honestly wasn't much new business." (Interview C4-1, Pos. 368–396)

The head of data considers his work to be a continuation of the methods and mechanics that existed in "offline" data work, albeit on a larger scale:

[126] See also 8.1, "Machine Learning" and "Natural Language Processing"

> "Research through household waste was before my time, but I was still familiar with [physical] index cards. No, I don't think it's a truly new thing. We used to have market research where you surveyed your subscribers for three months, paid twenty thousand Euros and then received a sample on the day of the survey deadline. And now we have all of this data on a daily basis, free of charge and in much larger samples. It's simply a development." (Interview C4-2, Pos. 563–570)

Attesting to this claim, the amount of data points and the size of data inside the publisher's data lake has increased exponentially since its inception:

> "We now have around seven thousand dimensions and metrics in our lake across all data sources. As a reference value, we started with 850 megabytes of pure text files from tracking data via Snowplow per day. Now we are at eighteen gigabytes per day. In other words, pure text files of website events only, which we process, analyze, aggregate, combine and then make available. And that requires a different infrastructure, which we are also constantly developing." (Interview C4-2, Pos. 179–189)

Having operated a regional distribution infrastructure, the publisher also had to gather and optimize logistics routing data early on, giving the company a head start in applying this data expertise elsewhere:

> "We are driven by logistics when it comes to optimizing routes or districts [...] for fifteen years. [...] Therefore, we recognized the signs of the times in terms of data-driven optimization earlier, it's been in our DNA for quite some time. We already had an inkling of what is now possible with customer data, but we didn't have the means." (Interview C4-2, Pos. 575–581)

In the future, lessons from digital assets may inform print production: “Incidentally, I believe that the more digital customers we have, the more we will be able to deduce decisions for the newspaper from digital habits in the near future. And we can do this systematically using algorithms.” (Interview C4-1, Pos. 714–717) Compared to the period before the establishment of the data unit, the head of data summarizes, data work and data orientation has become firmly embedded in upper management: “I think the biggest difference is that we now believe in data and that there is no alternative to our work. That aptly describes our mindset from a management perspective.” (Interview C4-2, Pos. 555–557)

Challenges and data objectives

As the data team entails high setup and operational costs, it plans to expand its headcount and achieve demonstrable returns on investment by 2025. (Interview C4-2, Pos. 584–586) Revenue could be generated through the development of new data products, while operational gains may be realized across multiple units or by offering data products to external customers: “We aim to measurably increase success across all areas. [We also aim to] expand the sale of services based on our expertise and the data we collect. [...] Perhaps even sell them to the outside world as a stand-alone product.” (Interview C4-2, Pos. 587–596) New products could be conceived around analytics software, internal bots as data delivery channels, or even pre-trained machine learning models: “In other words, models that we have developed and that are just waiting for us to feed in more data and then play out the results. Our model trains itself to get even better with more data and be even more precise.” (Interview C4-2, Pos. 603–613) Yet, the head of data envisions operational and supply chain optimization as the most significant revenue stream for his team. In his view, the data team resembles more of a management consultancy:

> "I don't think refinancing via external products is realistic. I'm also talking about internal optimization. When I think of the example of logistics and the cost apparatus and the number of people working there and rising minimum wages. We absolutely have to optimize there. We could definitely make our contribution, leverage data in connection with our printing presses, for example. If we read out and predict on machine data in order to prevent breakdowns or further malfunctions or damage." (Interview C4-2, Pos. 659–670)

Otherwise unapologetic about the extreme focus on data and performance measurement, the executive acknowledges a certain anxiety among employees: "Not everyone can manage this mind shift and this change. [...] Dealing with data triggers a fear of control in many people, so it's my obligation to take the people in the company along so that they perceive data as a helpful tool." (Interview C4-1, Pos. 644–658) To counter potential negative sentiments towards data, regular training sessions and workshops are held across all departments. More senior members of the sales team, the head of data states, tend to be harder to convince (Interview C4-2, Pos. 457–462). For this reason, knowledge transfer evolved into a main objective of the data team:

> "Every quarter there's a deep dive from a data science perspective. It definitely makes people's eyes widen a little. If you tell the lady or gentleman from accounting about your churn framework, without meaning to sound disrespectful, things fail due to a missing shared vocabulary. But concerning the basics, I try to have everyone tagging along and give them the opportunity to jump on the bandwagon." (Interview C4-2, Pos. 472–486)

People less inclined to tap into data remain in pivotal positions at the publisher. Here, the head of data sees another critical barrier to the data team's success: "There are some people, especially in editorial, who doubt the numbers. [...] If such people are in key positions, multipliers, then it becomes a little tougher. I also see that as a major hurdle." (Interview C4-2, Pos. 637–641)

On the problem of data interpretability, the interviewees provide technical cues but never reflect on their own agency or the power attached to data work. In slight contradiction to the statement above, the head of data recounts how the credibility of data is no longer questioned. Initially, data provided by the nascent data team had a validity rate of twenty percent, which the head of data claims has now grown to one-hundred percent: "We can't attribute the last two percent, so we just leave them out. Because what we can't explain, we don't show but improve things in the background. It's a slow process so that the credibility of the data is no longer questioned." (Interview C4-2, Pos. 351–360) In terms of technological challenges, the data scientist believes that the biggest hurdle for publishers lies in the data collection and preparation phase:

> "These days you don't have to understand everything, because [software] packages generate predictions without much effort. You don't have to understand what a *support vector machine*[127] is or understand the parameters—just do *hypertuning*[128] and it works. I think the biggest problem today is data preparation. [...] If clean data goes in at the front, then prediction is really no longer a challenge today. But the road to get there is really rocky." (Interview C4-3, Pos. 538–546)

[127] See also 8.1, "Support Vector Machine"
[128] See also 8.1, "Hypertuning"

Another goal expressed by the data team is to integrate data into various tools and applications along the news production chain, such as the CMS. This approach ensures that data continually informs processes, rather than being analyzed or viewed on an ad-hoc basis through dashboards:

> "We need to move more in the direction of product development—so that not only selective analyses are carried out, but [machine learning] models are made available for other systems. [...] So that when editors write they also have these tools available at their fingertips, within the CMS. They come across other topics, they receive suggestions, all this knowledge that we build up in the background is always directly available. Not just in the form of a dashboard, but integrated into day-to-day work. That's what we'll be working on in the coming years." (Interview C4-3, Pos. 558–567)

Another important lesson learned by the data team was the extent of explanation and training required for editors to effectively use dashboards. The preconceived notion of data self-serviceability was quickly shattered: "You hand over access to Tableau, give everyone a brief introduction and then they go about everything themselves. That was the idea. A complete failure." (Interview C4-2, Pos. 246–248) Once the highly transactional nature of their exchange with editorial became evident, the data team then implemented a project management approach to assess requirements:

> "That's why [...] whenever a new inquiry for a data science analysis or a new dashboard comes in, we start with a project canvas where we answer ten or twelve questions for ourselves and define very clearly: These are the expectations. That's the reason why we're doing this in the first place. This is the goal. These are the metrics. And then the other party signs off on it.

> Everyone has to deal with what's in scope, because otherwise we'll end up on a hamster wheel." (Interview C4-2, Pos. 258–266)

At the time of the interviews, these projects are far from fully automated, often taking several months to complete: "Our approach is to allow only one quarter to pass between the idea and the proof-of-concept, or go-live, of a first prototype." (Interview C4-2, Pos. 270–272)

Frequent inquiries about the status of the data system or the validity of specific data points quickly prompted the data team to provide "data about data" on the status of their data infrastructure through a website and knowledge base:

> "It started with personal messages, then groups in Microsoft Teams. At some point, we no longer had an overview of all the channels and the effort grew too high. We decided to change things again and launched our own website, which we also launched in the context [of restructuring our digital department]. On this website, we present our three teams [Technology, Data, and Digital] and also blog about new things we add." (Interview C4-2, Pos. 367–375)[129]

By prioritizing the delivery of local news, the publisher aims to establish a competitive advantage by cultivating an advertising ecosystem fueled by targeted data and comprehensive insights into the demographics and financial standing of local businesses:

[129] I have read the corporate blog in its entirety and confirm how the facts and statements given in the interviews are representative of the content published there. The specific URL of the corporate blog is not disclosed to avoid identification.

> "In the advertising market we see we have more information than a global player might have or be interested in. [...] Public data sources that can be enriched with commercial registry data and so on. We are currently tapping into these sources and forecasting the potential [of advertising customers] based on revenue and employee numbers. We go even further when it comes to crawling. That's what our data scientist does. Crawling yellow pages or crawling Facebook and simply looking at how people talk about certain products on Facebook. [...] So then, in a sales pitch, we actually know more about the potential customer than they do." (Interview C4-2, Pos. 422–441)

At the same time the demand from local advertisers for performance data on campaigns or native advertising pieces has only emerged recently, as of 2020. The publisher clearly finds itself in a market that lags behind the national markets targeted by larger publishers in the sample: "How successful is something I do? That is apparently not yet so pronounced in a local advertising environment. [Advertisers think] the (print) newspaper printed my article, people will read it. But I would want to be much more numbers-driven and data-driven." (Interview C4-2, Pos. 453–456)

Data privacy

Data privacy remains a technological and cultural hurdle for the publisher. On one hand, technically feasible data processes are stifled by legal barriers. Big platforms with large financial resources and a core business model less reliant on journalistic integrity could easily absorb punitive damages from privacy violations:

> "Data protection means trust. What we are doing, profiling, is a loophole in the legislation somewhere. [...] We could do a lot more, but we don't even

know if we are allowed to. For the sake of our brand, which stands for trust, we don't. This holds us back, where perhaps a Facebook wouldn't give it a second thought." (Interview C4-2, Pos. 623–629)

On the other hand, there is a need to educate people about the legal limitations and caveats surrounding data: "We also need to learn about culture in-house. We are currently running a learning program here at the company on how to handle data with all colleagues. I find this rather exhausting. It feels a bit like adult education [Volkshochschule] to me." (Interview C4-1, Pos. 608–609) With growing volume of data and data infrastructure, data security becomes increasingly vital for the publisher. This importance was underscored by a recent hacker attack on a national competitor:

"On the other hand, data security is incredibly important for a company like ours. [...] While total security is never achievable, but we can do everything in our power to avoid open flanks. What [competitor] experienced was highly damaging to business and, in this respect, an awareness of how data gets handled in a company that lives on digital content is super important." (Interview C4-1, Pos. 604–621)

Dashboards

Incorporating various dashboards into their daily routines, all editors on staff routinely review different dashboards (Interview C4-3, Pos. 380–385). Explaining the launch of a new dashboard that would enable editors to explore topic clusters with greater granularity, the data scientist reiterates his role as a gatekeeper of more complex data, assuming that editors would otherwise feel overwhelmed:

> "We test-drove this dashboard several times to get people used to it. Only the editorial development team receives the dashboard, not everyone else, because they would probably be overwhelmed. When a new technology like this comes around, we have to do a lot of training. You can't expect editors to understand everything right from the get-go." (Interview C4-3, Pos. 385–390)

While dashboards have become a standard interface for operational data and a device for tracking the publisher's chosen metrics for overall success, there are also critical statements about dashboards. For example, other interfaces or channels might offer more personalized and easy access: "These things are gimmicks, but our [Microsoft Teams] bot is a real success. [...] The long-term goal is for the editorial team to be able to get their information easily via several channels and not always have to go to a dashboard." (Interview C4-3, Pos. 392–396) Similarly, the head of data acknowledges that dashboards might not always be the optimal vehicle for conveying information, or they may be disproportionate to the data question at hand: "There's also Microsoft Excel for example. It's been around for a while and has its justification! When in doubt, a 'yes' or 'no' might be enough already." (Interview C4-2, Pos. 312–315) Another account touches on how dashboards are evolving into boundary objects, intended for various stakeholders across the organization to access their performance data and self-regulate:

> "We will soon be releasing this tool [for splitting customers into segments]. We are trying to address different people [inside the company] with the tool. Not just those who deal with churn, but so that everyone in the company can contribute to churn prevention and analyze whether their area plays a role." (Interview C4-3, Pos. 437–445)

Technology and tools

Essentially a best-of-breed strategy, we find a combination of different software parts from multiple cloud vendors in this case. After committing to a SQL-like query technology for large data sets called Presto, the publisher found the software worked best on Amazon Web Services (AWS): "The only solution we found that also deployed well comes from AWS. That's why our data lake is in AWS, because it allows us to talk to the data more directly." (Interview C4-3, Pos. 97–104) Other parts of the data technology stack are based on Google's BigQuery[130] with further migration toward Google's cloud offering planned, adding a third service provider to the mix.[131] Upper management allows such fragmentation across multiple cloud vendors as it appears to reflect an overall willingness to experiment: "We always said that we do not insist on a system for compliance reasons, instead we simply use what is best for each use case." (Interview C4-2, Pos. 192–197) As for analytics purposes, Google's widely adopted yet cost-prohibitive solution was quickly dismissed. Instead, the team built custom analytics based on the open-source product Snowplow: "With [Google Analytics] the free version wasn't enough because we need data in real time and have lots of events. Google charges six-figures. So then we implemented our own tracking [...] Snowplow is wonderful. We integrated it together with our in-house engineering team." (Interview C4-2, Pos. 102–109) Overall, the dynamic technological situation seems to match the ongoing experimentation with data and data infrastructure.

[130] See also 8.1, "BigQuery"

[131] "In the background, everything no longer runs via Presto but BigQuery because Google has developed very well in the direction of big data. That's why we're slowly moving all our processes in that direction." (Interview C4-3, Pos. 364-366)

Recently, the data team began automatically categorizing and clustering their news articles. Topics are constructed around keywords, referred to as tags, such as notable figures, locales, products, or sporting teams. The ultimate objective is to uncover new dependencies in a relationship graph illustrating the connections between these topics so that editors are able to identify high performing topic combinations. Additionally, the data scientist incorporates external references and topics into the graph, allowing editors to uncover gaps or opportunities in their news coverage:

> "We want to inform, but not be stuck inside our filter bubble. Which means that we also expand our offering to include topics that are not currently covered by us. [...] We have crawlers[132] that monitor Google, social media, data from various newspapers and recognize what gets reported. In this way, we enable our editors to stay up-to-date and discover the relevant topics for the day or the week. We apply text analysis and graph theory in the background. It's very scientific in nature." (Interview C4-3, Pos. 194–206)

Perhaps counterintuitively, the data team employs machine learning to predict churn for print subscriptions, but not for digital subscriptions or dashboards: "Using so-called SHAP[133] values, you can really see for each individual customer which parameters play a role for them. Customer A has a probability of ninety per cent because he is over sixty or over seventy. Customer B has the same probability because he issued a lot of complaints." (Interview C4-3, Pos. 473–498)

[132] See also 8.1, "Crawler"
[133] See also 8.1, "SHAP values"

6.2.5 The magazine publisher (C5)

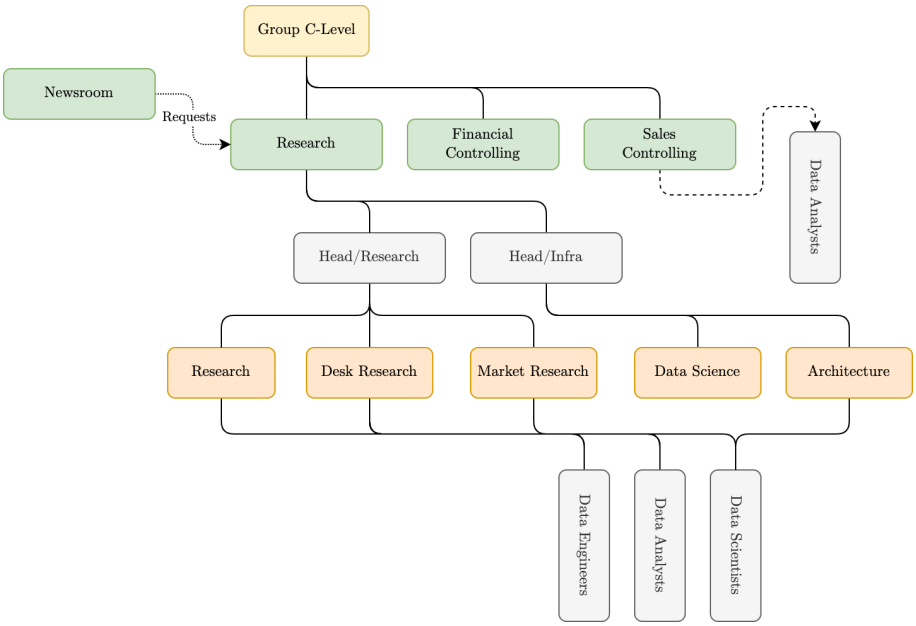

Fig. 5: Organizational chart of data work at C5

Organizational structure

In this case, the data and research team is organized along four pillars, as explained by the head of research: "There's research with primary research consisting of qualitative and quantitative, then we also have a section of secondary research, desk research. That's my strand." (Interview C5-1, Pos. 97–99) The adjacent data team covers data architecture and data science: "This mix then results in an individual analysis team, which comes together depending on the project." (Interview C5-1, Pos. 99–103)[134] Deviating from the organizational structure found in other cases, a cross-functional approach facilitates the convergence of various data-related roles around specific projects.

[134] Here, quantitative refers to social science methods like surveys and regression analysis, which are also used by the market research subdivision, a recent addition to the data and research team (Interview C5-1, Pos. 709–711).

While the quarterly steering panel discusses the agenda and prioritization of the data and research team, the team itself retains its interpretive autonomy:

> "[Our boss says], you know the subject matter better anyway. I just want the processes and structures to be right. In other words, we are more or less completely free when it comes to the specifics. [...] In the steering panel, we talk about our key topics on a quarterly basis. It's more about prioritization conflicts, because we can't work on as many issues as we receive requests." (Interview C5-1, Pos. 154–166)

While the data and research team has no clearly defined leadership role, the head of research informally shares the duty with a more technically oriented colleague: "In fact, it has grown in such a way that I very much take the lead in the task-specific part. It's more of a customary right and has to do with the historical fact that I was first." (Interview C5-1, Pos. 761–765)

The financial controller describes the publisher as decentralized. This explains why, at the group level, there has been little change in the mechanics and tools used for financial control, whereas other units use software from Salesforce, Microsoft, and more advanced analytics to do the task:

> "Operational control matters are handled by the sales department. Which kiosk sells, how or which area sells how much and possibly also why? This takes place in specialist departments, which certainly also use analytics and Salesforce. All I know about these software applications is their annual license fee and the name, to put it mildly." (Interview C5-4, Pos. 15–21)

In stark contrast to the depiction given by the head of research, where data and research are portrayed as the authority on all matters concerning quantitative and qualitative data, the engineering manager consistently emphasizes the narrative of decoupling, flat hierarchies and matrix management. Without mentioning the data and research team once, she describes data science as a fluid responsibility "where it longer plays a role where people are anchored" (Interview C5-2, Pos. 306–309) and which she considers atypical: "It's new for some of our traditional department heads, because people like to think in terms of having departments as large as possible." (Interview C5-2, Pos. 309–310)[135]

Actors and role perception

A total of four interviews were conducted, again via referrals according to the snowball procedure (Atkinson & Flint, 2001). A new set of roles emerges, perceived as crucial to data-related work within the organization by the interviewees: the head of research, the head of innovation, a financial controller (looking at financial metrics), and an engineering manager involved in developing the organization's infrastructure and software.

The financial controller aggregates profits and losses for the group, using accounting software by SAP. He describes his work as "data groundwork" (Interview C5-4, Pos. 298) involving repeated analysis, projection, and evaluation based on monthly cycles. "All of it neither modern nor particularly progressive. I think it was set up 20 years ago and our SAP still looks the same." (Interview C5-4, Pos. 3–8) As the interface between the control units,

[135] On its corporate blog, the organization reports on a novel project-oriented approach adjacent to matrix management where task forces are assembled for specific projects and then dispersed upon completion. In order to not identify the case here, I do not disclose the specific URL of the blog.

reports are passed on as static documents, highlighting the comparatively lower technological sophistication to other units handling data: “I receive a PDF from an upstream database and type it up. [...] These are all well-established structures that are eventually no longer up to date. That’s the nature of large organizations.” (Interview C5-4, Pos. 71–110) Explaining his duties further, the financial controller describes various balancing and reporting tasks:

> “Our task is to control and organize the cost centers so that revenues and costs are allocated correctly. So that at the end you can analyze whether an object has sold well or badly. [...] Furthermore, it is our task to prepare a monthly financial statement for the company, to see how things are going financially. So departmental analyses, comments, justifications. That’s our day-to-day business.” (Interview C5-4, Pos. 48–59)

With a background in journalism and academia, the head of research has worked at the publisher since the beginning of its data initiatives, operating under a new product leadership: “We’re so new, we don’t even have a proper job title. I’m just a researcher. I started there [<5 years ago] and basically helped develop [data & research] from its inception to its growth and current work.” (Interview C5-1, Pos. 46-49) He describes his work as project-oriented with little to no data automation or recurring tasks: “My work is strongly driven by primary research and analysis. [...] I don’t have a routine in the sense that I look at ten dashboards every day to check in on things. My work is more about realizing individual research projects.” (Interview C5-1, Pos. 425–430) At multiple points in the interview, he emphasizes how his expertise in the social sciences and methodological foundations are crucial to the data and research team’s work: “With my strong academic background, I have always brought a very academic and scientific approach to data work.” (Interview C5-

1, Pos. 72–77) Similarly, the head of innovation describes their role as consisting of facilitation, project planning, execution, and research into developing fields and topics (Interview C5-3, Pos. 8–19): "Data is very, very important to me. And I have to admit my soft spot for data due to my history in science. I like data and find it incredibly important to engage with it." (Interview C5-3, Pos. 32–36)

The engineering manager has a background in print publishing, takes responsibility for all web and app products at the publisher, leading teams of software developers, project managers, and designers. She emphasizes the fundamental importance of data to her work, considering herself "certainly not as someone who analyzes the data directly, but someone who uses data to implement things" (Interview C5-2, Pos. 21–23).

Origins and changes

Initially, data efforts were launched within the editorial team, with several editorial functions allocated to the task alongside their regular responsibilities. This attempt included the head of research, still a regular editorial analyst plus another more technical colleague, who "really worked on the [tracking] pixel basics, the data architecture in the back end. [...] The whole thing was relatively fruitless, it didn't work very well" (Interview C5-1, Pos. 56–64).

A second attempt was more successful and was initiated by the head of product, who installed data and research as direct reports (Interview C5-1, Pos. 71–76). A shifting focus from analytics to include reflection on data work and working with social science methods led to a name change:

> "[A few years ago] the department was first called analytics, then data analytics. And then I renamed it into [data and research] for reasons of our subject matter. For me, research means concentrating on questions. Calling the whole thing 'data' would be reductionist to me, because it completely ignores the fact that you also have to deal with the questions and not just with the methods. There's this data fixation in many publishing houses, where there are an incredible number of dashboards and an incredible amount of data used for control. But at the end of the day, you know relatively little about fundamental questions, what users want." (Interview C5-1, Pos. 79–91)

With a shift in organizational affiliation from editorial to product, the team moved away from merely providing data in the form of dashboards: "I believe this was a decisive development for us. Moving away from editorial work, from data work in the sense of providing a dashboard, to a department that is more interested in analyzing, understanding, and supporting." (Interview C5-1, Pos. 485–489) Operating in an organization with editorial dominance, several steps had to be taken to legitimize their new approach, from lobbying at the highest level to presenting their work in a more narrative way:

> "When you start giving fancy presentations and show the three issues that concern us today in a relatively structured way, then it has a completely different quality. Suddenly, you're in dialogue with the editors. [...] Suddenly, we were no longer just producing dashboards in secret, but we were recognizable as people who know things." (Interview C5-1, Pos. 493–506)

Throughout its existence, the research team built a solid reputation by presenting itself as a critical and academically-minded partner. The increased acceptance of data by the editorial team was contingent on this established role:

> "I think that has been one of the key developments of the last three years, that we have had this differentiation and this has also been accompanied by the increasing importance of research. [...] I've had regular meetings with the editor-in-chief and management, and the status of research and, therefore, also of data has risen dramatically [in the last few years]" (Interview C5-1, Pos. 116–119)

Now the teams finds itself in a situation of high visibility and data awareness, where expectations run exceedingly high: "We are getting into a position where we are a bit oversold. [...] We don't have a crystal ball that can make predictions." (Interview C5-1, Pos. 546–552)

Reflecting on the questions of the innovativeness or novelty of data work compared to the methods previously used in subscription marketing, the head of research believes, "that there has never been this level of insight into what our readers do and don't read. This represents a qualitative leap in user retention. [...] I think this is all completely new territory in the relationship between editors and readers." (Interview C5-1, Pos. 829–847) Taking a more nuanced view, the engineering manager sees a continuation in the use of statistical methods, albeit with much greater demand: "We always had a number of people who have been quite involved with opinion, opinion-forming, and statistics. Today, this would be summarized under data science, but the focus is slightly different. [...] We realize that such roles are naturally becoming

important." (Interview C5-2, Pos. 273–279) She envisions machine learning, and to an extent statistical learning, as tools enabling a comprehensive understanding of the business' inner workings, rather than solely as a manifestation of innovative audience measurements:

> "We still have plenty of information that we can tap into. The question is whether you call it data science now. I think that's a bit of a trend. You need to know how your business works, who your customers are, how customers behave, how you work, how your processes are, and that's important for every company." (Interview C5-2, Pos. 353–357)

Describing how data crosses boundaries, the engineering manager acknowledges that data today has many perspectives and meanings attached to it: "It starts with the data scientist trying to unearth some information deep down. But it can also simply be someone like an SEO or a technical SEO who says that the content of the HTML can be optimized to be found on Google better. [...] There are different perspectives on how we understand data today." (Interview C5-2, Pos. 469–478)

Looking into the future, the head of innovation highlights the boundaries that constrain what the publisher is able to do with data:

> "I know there's always an ethical component in [comparable organizations], a role model function. Because you can't constantly rant about Facebook and then do no better. [We are] purpose-driven. As a general trend, data work is becoming more important in terms of generation, processing and utilization. I don't think robot journalism will happen." (Interview C5-3, Pos. 417–439)

The researcher doubts the current pace of growth in his team will continue: "If we continue to grow at this rate then in five years we'll be at 50 people. I don't think that's going to happen." (Interview C5-1, Pos. 911–913) While the team currently reports to the head of product, the association might shift from product to sales (after moving from editorial to product before) in the near future: "I actually think it would be best if we as a department maintained this independence, because it naturally gives us the opportunity to represent the view from the outside much more credibly." (Interview C5-1, Pos. 921–933) Overall, the researcher sees the agency of individuals like the head of product as the main factor in the further development of data and research: "As with so many new things, this depends very much on individuals. And that's what makes it so difficult to predict." (Interview C5-1, Pos. 949–950)

Should we view data work at digital news publishers as an inherently new phenomenon or as a continuation? As the head of innovation points out, we have to differentiate between media, with the television industry closely tracking audience behavior way before the advent of the internet:

> "With television, ratings were the most important thing. [After broadcasting] you looked where the jumping-off points were. Where did the viewers go? Did they come back after the advert break? [There were] an incredible number of tests where people sat in rooms and watched the program. [...] In private television, there was always this sovereignty of data going on." (Interview C5-3, Pos. 197–204)

Meanwhile, the news industry remained complacent towards data, they continue: "In the eighties, when newspapers were money-printing machines, data work was less important. The data was there, but why look at it if it

didn't matter what you were doing?" (Interview C5-3, Pos. 216–228) To the engineering manager, current data work represents an evolutionary process: "Tools have got better; the amount of data has certainly increased. Yet the questions have remained the same, you always wanted to know your customer, [...] how they use the product, whether they're willing to buy." (Interview C5-2, Pos. 649–654)

Another relevant, inter-organizational development lies in how news organizations have begun to exchange information and cooperate. Here, the publisher openly exchanges information with competitors tackling the same general problems, which would have been unthinkable previously: "There is much more communication between publishers than before. Because we're all in the same boat. [...] We share information, tell others what works well or less well for us. That helps other people not to make the same mistakes." (Interview C5-2, Pos. 605–615)

Ongoing data work

As an internal customer of the data and research team, the head of innovation describes the trifecta of data work in more detail. Testing the waters for a new educational product, they first relied on secondary analysis using market research data, then they conducted qualitative interviews, and finally tracked success and acquisition metrics in the public product:

> "We looked at market media studies to gauge what the level of interest is like. These market media studies ask about everything, from favorite sites to the last issue [of the product] that people read or their level of income. [...] We then started to launch test balloons on this basis. Courses that were free initially. We then spoke to the users of these courses in person to bring

in this qualitative perspective. We launched the project and are now, of course, in the process of measuring its success and looking at where the actual buyers come from." (Interview C5-3, Pos. 66–79)

At the publisher, a significant emphasis is placed on automation based on metadata, meaning data that describes other data—articles in this case. Metadata has the potential to provide readers with improved orientation:

"We actually have a few projects where we are thinking about how we can use artificial intelligence in our data work. How can we extract metadata from existing archive material, either for research or to make it more useful for our users? We found out that many people want perspectives, different angles to a topic. This led us to the question, is it possible to simply extract the perspective on a topic from metadata? But at the moment, the metadata just isn't good enough for that. And you can't make someone type it all in. The question is whether sentiment analysis will make it easy to do this." (Interview C5-3, Pos. 399–409)

Another internal project involves automated tagging across the article database, even extending to content from competing publications: "[In documentation] there is a department for competitor monitoring, where thousands of articles are indexed and fed into a database. They have this categorization system that perhaps might be a little antediluvian from a data analysis perspective. This is also being revised." (Interview C5-3, Pos. 410–411) While machine learning assists in enriching archival data with metadata, selling raw data as training data for third-party models is not on the agenda for the publisher (Interview C5-2, Pos. 503–511). However, according to the engineering manager, statistical learning methods could potentially unlock new

markets. Automating the generation of natural language from data and translating natural language into data points, the publisher currently runs various text-as-data experiments:

> "As a new idea, we are currently working with start-ups on how we can better understand text. Can we better cluster text information automatically if we understand whether there are positive or negative statements in a text on a certain topic? [...] There's a lot we can do with data to support traditional businesses. This may also result in new products. Things that perhaps also bypass traditional journalism a little, are off-topic, but don't compete with it either." (Interview C5-2, Pos. 689–704)

Privacy issues

In compliance with GDPR regulations[136], the publisher segregates individually-identifiable customer data from tracking data. With data shared among different business units, there are limitations imposed as to what the publisher can do with data, as the publisher distinguishes between anonymoued data and data used for customer support: "If someone buys something, we know who they are, and if they have a problem, we can help. However, this information is not merged with general tracking information. There are data protection regulations which prohibit us from combining certain data." (Interview C5-2, Pos. 398–403)

Another significant challenge for the publisher is that GDPR involves consent management and ultimately leads to some form of degradation in user experience:

[136] See also 8.1, "GDPR"

> "I think the biggest discussion we are currently having is the whole area of third-party cookies and data privacy. I think that's the biggest area that all companies are currently working on. And we're only involved because it's ultimately the advertising industry that's addressed here. [...] I can't imagine having to click away fifteen consents and terms of service every time I visit a website. But that's what it could lead to if certain regulations are strengthened in this way. [...] Ultimately, you can do what Netflix does. [...] Only if users have signed up and registered, accepted certain terms, can they get any further at all. That would be the ultimate way. Does that help the information society? I'm not sure." (Interview C5-2, Pos. 733–754)

External factors

Drawing a key parallel to social media platforms like Facebook, the head of research highlights the pursuit of user authentication, which has evolved into a critical objective for publishers as well. By convincing users to register and log on to their digital products, the authenticating party gains the ability to track and enrich user profiles, optimizing their products:

> "On the other hand, there is certainly an impetus from ecommerce, a major pioneering force in terms of analytics. [...] In sales, there are also people who are working on optimizing the payment process. At the same time, there are different expectations around the product than in ecommerce. Meaning, the methods cannot be adopted one-to-one." (Interview C5-1, Pos. 857–878)

With respect to personalization, the degree to which content is ordered or arranged according to a user's individual interests or behavioral data, other subscription businesses' methods and practices should only be replicated to a certain extent:

> "Even though I know that you had preferred to watch cross-country skiing in the last few weeks, and you would get cross-country skiing recommendations on Netflix or cycling, then of course that's not what we would primarily present to you at the top of our website. Even if we know you're into it. Because we still think that the overarching news situation is more important to you." (Interview C5-2, Pos. 534–540)

Decision-making with data

Similar to the other cases, the data team is tasked with building a deeper understanding of the dynamics and success factors of the digital subscription business model. A process started by management and editorial, "who said that it would make sense for [data and research] to get together with the editorial team and report on what they actually know about users." (Interview C5-1, Pos. 183–210) On an individual basis for each department, the data team now faces the challenge of suggesting "what should and should not be paid content in said department. We went into these departmental discussions with theses, looked at the data, made initial analyses and said that they were the five points that came out for your department." (Interview C5-1, Pos. 185–210) Giving detailed accounts of the hypotheses provided in these meetings, the researcher's understanding of qualitative data work becomes apparent:

> "In the end, we conducted customized surveys for each rubric [...] And if you take them all together, it emerges that the texts that do well in culture are actually always those that are based around [our core brand]. Theatre reviews that have a political or social reference or book reviews that were political in some way do well. [...] This learning can then be applied to culture, but also to other departments that are more distant [from the core brand] such as sports." (Interview C5-1, Pos. 229–265)

Due to the decentralized structure of the group, issues arise in the delivery of data and metrics across units. Not only are data delivered in a static fashion, as email attachments, harmonization of data points towards individual requirements needs manual labor:

> "There really isn't a closed system in which things run from A to B and then a key indicator is drawn directly from it. That's a shame, but that's the way it is for now. I think it will stay that way. Each department has built its own solitaires." (Interview C5-4, Pos. 120–125)

Even internal access to "raw data" within the data team remains problematic, which in turn stalls big data analysis efforts: "[We've had] data scientists for [a few] years now. But they can't work properly because we don't have proper access to raw data, at the moment or in the past. That's changing now. It's all still very much in the making." (Interview C5-1, Pos. 738–741) Churn prediction through machine learning, an obvious concern of the data and research team, is addressed elsewhere: "One person does churn prediction in sales, but they're pretty much on their own. Like I said, at [organization] there is quite a structural cluster." (Interview C5-1, Pos. 782–787) However, the domain knowledge created there will not easily spill over to the data team either: "It's being broken up a bit in recent years, but still there. This also means that some of the things we do here are still happening in other departments, too." (Interview C5-1, Pos. 787–789)

Qualitative data, the head of innovation argues, should be gathered firsthand by the teams making product decisions based on the data. This approach aims to prevent the loss of details during translation:

> "If you do empathy work or super qualitative methods, then the [software] developers or the team themselves should be conducting interviews or engaging in participant observation. And no market research department translates it into a PowerPoint, censors things and produces a relatively anonymized presentation. You have to do some convincing for people to move this close to the actual data collection. Some argue this is not representative. And it's not supposed to be! [...] In the innovation process, I am in favor of having immersion in the data. The classic example: when you know the question within a quantitative survey, then you know much more about the answers than through a summary with thirty per cent objectivity written above it." (Interview C5-3, Pos. 281–320)

As the only interviewee in the sample, the head of research questions how data is generated in the first place and how it not only appears to be contingent on the measurability of any given phenomenon, but also on the decision-making power of data architects. These conditions, in turn, imply a variable amount of omission:

> "Tracking concepts basically already define what is measured and what is not. [...] This is noteworthy, because it has implications for later analysis. We can only measure behavior that produces data. Scrolling on our homepage, which is a totally crucial usage scenario, cannot be measured because there's no data about it. Why? Because it's difficult for data architects to measure." (Interview C5-1, Pos. 356–370)

After years of lobbying for increased data awareness, the head of research now finds himself managing excessive expectations and trust around data-informed decisions, at times advocating for journalistic intuition over data:

> "It's nice that you're noticing us now. It's good that you're looking at the users now, but don't switch off your journalistic gut feeling. We don't want to deterministically say what the editorial team should do. Our challenge right now is more of an educational nature, to show how you can actually work well with data. Simply because I am also a [reader of the medium] I don't want an editorial team that blindly follows my numbers." (Interview C5-1, Pos. 556–569)

The engineering manager emphasizes overarching journalistic values, which should always take precedence over data, even if the data suggests going against intuition:

> "I believe data will become increasingly important and will influence a large part of how we define, build and present our products and present information. However, I think we will also keep reminding ourselves that data doesn't necessarily mean you have to do it that way. We'll always have a journalistic ambition to report on what we believe is important and right." (Interview C5-2, Pos. 666–680)

Data and the newsroom

The researcher describes a reciprocal misunderstanding regarding the perceived editorial power shared by both readers and editors. First, the hugely influential enabling structures within publishing houses are often underestimated. In the past, this led decision makers to harbor excessive expectations around editorial innovation and a "positive format blindness" about how younger target groups could be reached by a new format:

> "From the user's point of view, everything that takes place at [website of organization] seems to be the work of the editorial team. The fact that hundreds of product managers and other people are involved is not perceived in this way, but for [the user] everything is just that, the mirror of editorial. In reality, however, it's a much broader area and a much broader set of expertise that's required." (Interview C5-1, Pos. 23–38)

Strategic decisions are now shifting to product managers or other roles within the publishing domain. An illustration of this transition is a new online publication targeting millennials, which ultimately failed because it was based on intuitive editorial assumptions:

> "Among other reasons, one reason [for termination] was that we were researching [vertically]. We had a sample of usage and behavioral data from under 30-year-olds. And we compared this with older users. It turned out that reporting on political parties was used more by young people than by older people. [...] What the editorial team would see as a dry, boring topic was totally appreciated by our young users. [...] The expectations of a target group are not primarily based on a format, but on information." (Interview C5-1, Pos. 578–618)

Directly validating one of the guiding assumptions, the researcher acknowledges that structural changes are driven by enablers on the publishing side rather than by editorial innovation: "In this respect, it makes total sense [to focus] on the publishing house, because the structures of journalism are controlled more by the people in the publishing house than by the editorial team" (Interview C5-1, Pos. 114–117) The publisher both plans and maintains the technological infrastructure for data work, with editorial sidelined: "You

can't do it without the publisher and you know that the people there have a pretty big influence." (Interview C5-3, Pos. 180–182) Several key people from editorial are said to be limited to structuring the editorial process itself. (Interview C5-3, Pos. 184)

On the other hand, editorial decisions can, and should, at times override clear paths towards increased revenue as outlined by, for instance, conversion metrics: "We know that advice and service pieces convert particularly well. Of course, we don't just do advice pieces, as it is not the aim of our editorial team to tell you where to find cheap radiators." Articles that tend to convert well are not the same as stories that are read in breadth and depth by subscribers, a learning from experimentation with the article index metric (Interview C5-1, Pos. 623–641). In some cases, gut feeling can and should override data indicating a decision-path, which the head of innovation defines as data informedness:

> "My opinion is, one should collect [data] and then be able to make a gut decision against it. [...] I don't think you have to be data-driven, but data-informed. [...] [In one case] on gender-inclusive language, a survey showed that it bothers people. At the same time, the journalistic perspective is to write in a gender-appropriate way. Because there is also an educational or social aspect to the work." (Interview C5-3, Pos. 137–166)

In their approach to data, the head of research observes editorial oscillating between resistance and outright embrace, viewing data as the ultimate truth: "Handling data, how it must be interpreted, how data is not always representative of truth—this may sound relatively basic. But [the editorial team] sometimes wavers back and forth there." (Interview C5-1, Pos. 128–132)

In stark contrast, other editorial staff reject data authority: "It goes without saying that our authority is not accepted by everyone. It does eat away at the self-image of one editor or another. [...] Probably also due to the fact that we are backed from quite high up. Absolutely a pain point." (Interview C5-1, Pos. 999–1007) Overall, successful data work at the publisher entails mediating between the extremes of editorial intuition and data-driven decision making.

By contrast to previous statements, the engineering manager asserts a significant editorial primacy over all product and data decisions (Interview C5-2, Pos. 173–178). However, it is worth noting that the head of product both established and oversaw the data and research team. As an indication of editorial agenda setting power, the engineering manager highlights the use of the management framework OKR[137]:

> "We organize ourselves on a quarterly basis and people from the editorial, product, and technical departments sit together. And we plan together what's coming up. In other words, I find out what the editorial team would like. I find out what moves the product, and, from a technical point of view, I say what things are on our plate in the direction we should develop." (Interview C5-2, Pos. 186–191)

In acknowledging multiple prevailing perspectives on the same data, the engineering manager touches on the concept of data artifacts as boundary objects: "[It] can be the same data, but there are two different perspectives on it or different perspectives on it. And we have that in many areas here." (Interview C5-2, Pos. 62–77) At the same time, she suggests a dissolution of department demarcations:

[137] See also 8.1, "Objectives and Key Results (OKR)".

> "That's the goal behind the introduction of OKR, that across different areas we define something like a funnel team. Meaning we have different tasks and team members coming together from different areas of the company. And that's why there is no longer a clear separation in the departments." (Interview C5-2, Pos. 233–241)

Another challenge lies in conflicting velocities between the data and editorial teams. In the past, organizational processes were adapted to the editorial news cycle. This collides with a data team interested in long-running experiments:

> "We conduct surveys with six thousand people. Add in data analysis of user behavior, we did it all in two months. But we are still slow by [the organizations'] standards. Large parts of the organization work at very short notice. Most clearly in the case of the editorial team." (Interview C5-1, Pos. 1017–1019)

Yet, critical phenomena such as churn might be better understood through large datasets and a long-term perspective. Especially in this case, the impact of churn is substantiated by absolute numbers. While the publisher gained hundreds of thousands of new subscribers year-over-year, the loss or churn also ranges in the hundreds of thousands:

> "Sales is at least quarterly driven, tends to have a monthly perspective. So there's a certain lack of sustainability from this short-term view. [...] We find bringing our perspective to the table difficult because everything is characterized by economic pressure and sales naturally wants short-term, quick success." (Interview C5-1, Pos. 1020–1046)

Metrics and data sources

Due to the data and research team's focus on understanding and optimizing digital subscription dynamics, well-established conversion and engagement metrics play a key role in its operation. More specifically, they recently developed a custom variation of a composite article score, "a metric that tries to contextualize article success. So instead of looking at an absolute number of reach, you look at a relative number that considers time of day, day of week et cetera. That's the quantitative aspect." (Interview C5-1, Pos. 276–282) Quantitative signals feeding into the article score are then qualified further by engagement metrics. But the publisher also ran into the same problem around composite metrics that others in my sample encountered:

> "Reading depth ultimately expresses a qualitative usage. But also another example of a data fetish! There are so many [comparable] indices that have twenty or even thirty variables in them. In the end, you no longer know what [these article indices] actually say. We have tried to break this down as much as possible. [...] How do we operationalize whether an article is read intensively? Through a form of reading depth." (Interview C5-1, Pos. 284–293)

As described in the organizational structure subsection above, the team also employs qualitative social science methods, acting like an in-house market research unit. From the researcher's perspective, coded qualitative data resulting from surveys and interviews also count as metrics. In his words, these constitute non-automated metrics: "Data that we get from surveys is also a form of metrics. And I would say qualitative interviews produce data." (Interview C5-1, Pos. 301–305)

In addition to first-party analytics data acquired through pixelation[138], which are stored and accessed via Adobe Analytics, public industry metrics such as AGOF and IVW also play a key role in performance evaluation. Sales uses another platform with a closed set of metrics, which remains inaccessible to the research team: "It's called Calypso. But I'm really out of it. What exactly it does and what it doesn't do. Then we try to get access to it. But it's not easy either." (Interview C5-1, Pos. 386–388)[139] A market research data provider, Best for Planning, is used extensively in advertising and helps in gauging interest for new products: "Consumer data is also included there. We simply looked at target groups or demographics to see how much interest there is in [certain products]." (Interview C5-3, Pos. 92–94) Overall, the publisher aims to streamline and harmonize the number of data sources, sometimes at the expense of editorial control over tools, aiming to narrow down the set of sources and even switching off some metrics, for example from the provider Parse.ly, which was used to optimize the home page (Interview C5-1, Pos. 378–384).

Dashboards

Despite the growing reputation and relevance of the data and research team, a recent editorial dashboard was planned and managed by an editorial developer, with the data team overseeing the technological aspects (Interview C5-1, Pos. 407–411). Dashboards here seem to primarily concern the editorial department, with the data and research team not managing any sales, marketing, or systems dashboards—which they regard as a superficial means of accessing data:

[138] Translated from the German expression *Verpixelung*, meaning the use of tracking pixels.
[139] Calypso is a software application aimed at financial institutions, promising to "consolidate their infrastructure on a single platform". In June 2023, the software effectively became the property of Nasdaq (Gara, Hughes & Mancini, 2023).

"Dashboards are used more in the editorial department. In the case of managing editors, they just look at the figures without being able to analyze them. Whereas our endeavor is to also understand why the number is going down or up." (Interview C5-1, Pos. 436–440) This lack of involvement with dashboards, a crucial management device in other cases, seems to be connected to a disregard for the technology overall. From the perspective of the head of research, dashboards provide limited utility and could even be dispensed with entirely: "The [editorial team] sometimes seems reliant on dashboards for live control, whereas we in the research team say switch them off for a week and not much will happen. You have that in your blood somehow." (Interview C5-1, Pos. 463–467)

Technology and tools

We find a similar data software assemblage as in other cases, with Adobe Analytics and Microsoft Power BI providing the building blocks for static reporting sent out by the research team. (Interview C5-1, Pos. 336–339) In addition to these popular tools, the data and research team makes use of academic software for qualitative analysis:

> "In addition to these survey tools themselves, here we use Questback for example, there is Excel Starter. Adobe Analytics could soon become more important. The same applies to qualitative tools such as MaxQDA, which we are currently testing. University research is an inspiration for us." (Interview C5-1, Pos. 391–398)

Reflecting on the relative inertia of his own unit, the controller identifies vendor lock-in with SAP as a major obstacle to innovation: "Once you introduce SAP, you won't change that. That's hara-kiri. SAP was introduced by [organization]

in 1994 and not much has changed since then, apart from regular updates from [SAP headquarters in] Walldorf." (Interview C5-4, Pos. 138–144) Although SAP appears to prohibit more dynamic processes, it functions as a form of data warehouse across the different controlling units:

> "SAP is really only there to find the data points, the data that you want to process further. Then you dump it into Excel or Power BI, do your analyses via Pivot or Excel and then you have the flexibility. SAP doesn't have any of that. SAP is a big tanker that has everything, but can do very little processing." (Interview C5-4, Pos. 222–228)

6.2.6 The national weekly (C6)

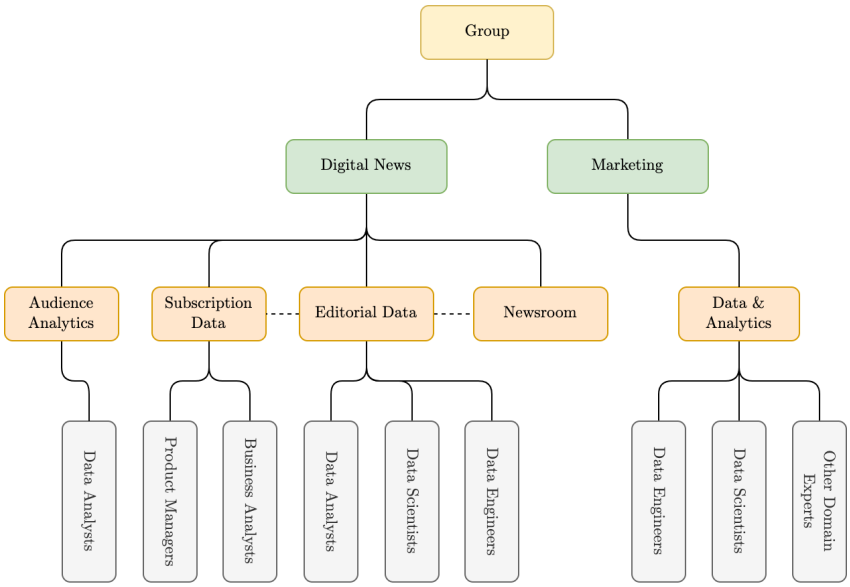

Fig. 6: Organizational chart of data work at C6

Organizational structure

With two dedicated data teams at the news organization (plus one in marketing at the group-level), data concerns are relatively centralized in this case. The subscription data team operates in a matrix organizational structure, aligned with three major sources of revenue, consisting of digital subscriptions, advertising sales for the advertising market and classifieds, each with its own director in the matrix: "The subscription data team works horizontally for all three business units plus [it also does the] delivery and technical measurement for editorial needs and reports." (Interview C6-4, Pos. 468–475) In said cross-cutting team, there are data analysts and data engineers, as well as service employees and most recently, business analysts, whose role it is to "moderate between engineering and stakeholders" (Interview C6-4, Pos. 480–492).

Service employees operate in similar ways to conventional service workers, providing information, generating reports, and responding to ad hoc requests through a ticket system. (Interview C6-4, Pos. 486–489)

Alongside the subscription data team, a dedicated editorial data team establishes the technological groundwork for data collection and storage in the pursuit of the mission "to become the single point of truth for everything relating to the company's data" while also ensuring to provide "key figures and data for a business area, process them and make them usable at the same time" (Interview C6-2, Pos. 62–66). In the data refinement cycle, before data is reconfigured to address specific inquiries and transformed into visually appealing "data products", the digital data team must first assess all the necessary sources, placing the team at the very beginning of the cycle:

> "We ensure the set-up and operation of a data warehouse. At the very lowest level, this means that we are responsible for tapping into a wide variety of data sources and also for solving all the technical issues. [In addition], we are responsible for data transformation into data that we can make usable in the form of reports and analyses or as the basis for our model calculations and, in the long term, in the form of data products that can then be used directly for personalization." (Interview C6-2, Pos. 33–41)

The editorial data team consists of data engineers, data scientists, and data analysts. Recently, a new role has been introduced to complement the original data engineer—a very specific position called DevOps.[140] (Interview C6-2, Pos. 521–522) With data engineering providing a unified technological basis to operate on, data scientists and data analysts collaborate cross-functionally and

[140] See also 8.1, "DevOps"

interdisciplinary, alongside sparring partners from product management and the editorial department (Interview C6-2, Pos. 517). To bridge the gap between engineering and science or analysis, the editorial head of data intends to establish a new role solely focused on requirements engineering[141] for the publisher: "[This] will ultimately establish a strong triangle between data engineering on our side, engineering in IT, and stakeholders. [...] taking on the topics of requirements management, what should reporting standards look like? How are the company's requirements changing?" (Interview C6-2, Pos. 524–546) To prevent data work for the sake of data work and ensure analyses make sense economically, cross-functional teams are used to keep both ends in check at the publisher: "An analyst always needs a sparring partner from the business side." (Interview C6-2, Pos. 795–797)

Among the revenue-streams mentioned earlier, data holds the utmost importance for paid products. As a result, the publisher consolidates all customer relationships and activities related to marketing, managing, and selling digital paid products under a single umbrella. This "paid" category can mean anything from subscriptions and market research datasets, to access to corpora of news articles. Two major columns make up this unit, customer service and product management. Still, the head of subscriptions asserts that there are flat hierarchies within the organization, describing it as "very structured by non-structure". (Interview C6-4, Pos. 66) Internal customers vary, and the data team primarily mediates to get to a deeper understanding of data in general: "The people we work with come from very different areas and are not so much attached to an organizational chart. Rather, the challenge is to work in an interdisciplinary way with colleagues from the editorial team. Specialized roles have developed." (Interview C6-4, Pos. 71–79)

[141] See also 8.1, "Requirements Engineering"

Increasingly, the publisher leans towards dynamic and temporary teams assembled to address specific problems. Here, data workers are relegated to adding expert input to these teams, alongside representatives from other technical departments:

> "We are working less and less on a departmental basis, but rather set up project teams with people from content management and paid product management or from ad sales or even from product development with the editorial team. A representative from data, for example, then sits in with a representative from [audience analytics] and representatives from the other departments and they discuss the topics together based on their expertise." (Interview C6-1, Pos. 40–45)

In managing these moving parts, the data team relies on human judgement to determine if the whole system still follows a certain internal logic and consistency: "If the data team doesn't know what's going on in the company, then something has definitely gone wrong. [...] we look at the most important figures in the most important areas every morning and gauge whether it all still makes sense." (Interview C6-2, Pos. 354–356) These data consistency observations are mandatory, as whoever reports on data at the publisher also bears responsibility for the accuracy and veracity of that data: "We are the owner of the figures that are reported. [...] Let's suppose we were to see in a report that reach is collapsing, our first task would be to see if there's something wrong with the data." (Interview C6-2, Pos. 377–405) Other tasks include monitoring legal compliance in gathering sensitive data and safeguarding it against security breaches (Interview C6-2, Pos. 70–72). However, to non-editorial stakeholders, the editorial data team's mission is not as clear-cut as its name implies—as we will see.

One major framework for data work revolves around the customer lifecycle, where the audience analytics team assumes ownership over the point of first contact with users:

> "If you look at it from a subscription perspective, my team is at the top of the funnel. At the publishing house, we are organized into product departments, which are responsible for the products, and line departments, which tend to provide infrastructure and work with people. [Audience analytics] and data are line departments for us. A lot of our work is communication and support." (Interview C6-1, Pos. 46–51)

As a side effect of the professionalization of data work, autonomy in technical decisions has shifted from audience analytics to the editorial data department:

> "In the past, when we were still [previous audience-related unit], I would have proposed how I wanted to track ["verpixeln"] something to development. Today, I go to [editorial data team] and tell them what information I need and make suggestions. [...] Then they check it, make counter-suggestions in case of doubt or take the requirement into the development process." (Interview C6-1, Pos. 46–51)

At the group level, another data and analytics team addresses more operational data tasks. The team reports to and receives impulses from marketing, which the group head of data considers to be a "speciality" as compared to other publishers: "We are part of marketing, we are not a separate department. [...] Normally, these data teams are external. We are part of all marketing coordination processes. That is our primary stakeholder and I estimate that 80% of the tasks come from marketing." (Interview C6-3, Pos. 196–207)

The team consists of data analysts and data managers, although the latter's role is not clearly defined, as it predates the current group-level leadership. (Interview C6-3, Pos. 632–636) In addition to data analysts, freelance software engineers help with data preparation, a task the group head of data sees as distinctly separate from the subsequent analysis:

> "Data analysis, and what I call data preparation, are very different profiles in the team. These are different skills and very different people who do these two things. [In-house] we only do light processing [...]. For management and data engineering, we buy in external services. Data engineering means extracting, transforming, and loading data. This is the classic ETL[142] route. Data must be represented in a certain way and form so that dashboards can be built on it." (Interview C6-3, Pos. 825–850)

The group head of data sees his work as comparable to the journalistic production process, in his view, a passive arbitration of data: "What we do is actually not that much different, on an abstract level, from the [work of] our editorial colleagues. We get data, we refine it and forward it somewhere. [...] Internal users are both machines and people who receive data back from us with interfaces." (Interview C6-3, Pos. 48–54) More specifically, the group-level data team pursues three main objectives. First, measuring the success of marketing activities. Second, enriching and acting on legacy customer data through building predictive models and affinity scores (Interview C6-3, Pos. 70–71) Third, the team carries out "testing", but not on-site, as this falls within the domain of the other data team. While the first two objectives seem relatively clear, the latter remains somewhat ambiguous.

[142] See also 8.1, "Extract, Transform, Load (ETL)"

Actors and role perception

As we are dealing with a larger organization, the case interviews included two heads of data, one from the publisher side and one connected to editorial. Additionally, interviews were conducted with the head of subscriptions and an audience analyst, both managing a group of data workers. The head of subscriptions has held various product management roles at the publisher for several years, serving in the current position for >5 years. With statistical knowledge gained through management education, she feels better equipped to understand the recently evolved technicalities and details of his job: "[My statistical knowledge] is totally paying off because my job today is very, very much characterized by working with numbers and with colleagues who are much better at it." (Interview C6-4, Pos. 24–29)

A computer scientist by education, the editorial head of data spent several years in various non-journalistic companies working in the field of data warehousing and business intelligence. After having looked into the areas of data science and machine learning, she then intentionally focused her career path on data: "I realized relatively quickly that I didn't want to work purely as a programmer, but rather on the business side of companies." (Interview C6-2, Pos. 13–17) In addition to matters of hiring and staff management, she describes her role as planning and maintaining the data warehouse—a large, centralized data software system that contains heterogenous sources and makes these available to stakeholders. (Interview C6-2, Pos. 33–41)

The group head of data has a background in economics and market research. After comfortably working at the interface between pure data preparation and (Interview C6-3, Pos. 97–103), a consulting assignment turned into the current role at the publisher. As a consultant, he exclusively worked on data problems,

"from preparation to usage, the entire chain" and built predictive models as a data scientist in the area of custom analytics and financial analytics (Interview C6-3, Pos. 108–112). Asked about the purpose of his work, the group head of data describes his role as relaying data rather than interpreting it: "We [work] with business transactions, we aggregate them and somehow make them tangible. We don't do anything beyond that." (Interview C6-3, Pos. 122–124) Reflecting on his decision-making power, the group head of data acknowledges his teams' significant influence in shaping the initial questions that serve as the starting points for data analyses: "Our involvement starts early, it starts with defining the question. [...] What do you want with the data? [...] It used to be called requirements management or engineering." (Interview C6-3, Pos. 536–541)[143] The degree of influence over framing analyses depends on the level of technical understanding of the requesting party:

> "Within marketing [...] there are colleagues who actually discuss with us nearly at SQL[144] level and say how they need the select [statement] and also write pseudo code.[145] [There are] others who are very, very far removed from this, who come more from a creative background, but who also have a certain amount of trust in what we are doing." (Interview C6-3, Pos. 545–553)

With a background in search-engine optimization (SEO), the audience analyst has been with the publisher for <10 years.

---

[143] See also 8.1, "Requirements engineering"

[144] SQL stands for Structured Query Language, a programming language developed in the 1970s by IBM researchers Raymond Boyce and Donald Chamberlin to manage and manipulate relational databases. SQL provides a standardized way to interact with databases and has become the de-facto standard of interrogating large datasets. See also 8.1, "SQL".

[145] See also 8.1, "Pseudocode"

While her set of day-to-day tasks has changed significantly during her tenure, her team now is tasked with monitoring the overall distribution of the publishers' digital products (Interview C6-1, Pos. 35–37). In their day-to-day tasks, the head of audience analytics also assesses how interest, demand, and supply around search keywords drive traffic towards the publisher's digital offerings—typical SEO responsibilities. (Interview C6-1, Pos. 120–124)

Origins and changes

Initially, data work at this publisher meant working with web analytics to increase the reach of digital products and, in turn, boost advertising sales. As the individuals most closely involved with data and statistics, search engine technicians, hired to enhance and comprehend traffic from search engines, were also assigned to web analytics:

> "Many publishers hired SEOs back then because search was and still remains the strongest channel. These people then simply took on web analytics because they needed it for their work and so they had the job. [...] Over time, the whole data thing grew up. [...] People wanted to know more about what users actually click on and what the consequences are when we send them to certain pages." (Interview C6-1, Pos. 250–264)

Initially (5 years prior), management of the publisher's online division set a strategic focus on pursuing data, improving data infrastructure, and enhancing data competence. Individual actors on the editorial side also expressed demand for better data: "The pressure kept increasing from all sides." (Interview C6-4, Pos. 457) This increased pressure to deliver data-informed insights can be attributed to the business model around digital subscriptions, where each subscription sold has a direct impact on revenue:

> "In our publishing house, closed content and subscriptions have led to many stakeholders in the upper echelons, who naturally have to take a bird's eye view of things, developing a desire for data. [...] Digital subscriptions have significantly accelerated this development because the link to sales is clearer." (Interview C6-3, Pos. 221–227)

On an organizational level, the increased emphasis on data work can be traced back to <2015, when the publisher started a project in the area of customer relationship management (CRM) software. The core data team at the group level was originally hired in the context of this project (Interview C6-3, Pos. 151–155). By the time the group head of data joined the project in <2018, it had already progressed beyond the architectural stage of setting up a data warehouse and had shifted its goals to predictive modeling:

> "The aim was to truly build predictive models on [the basis of the data warehouse] and create training data sets at the individual customer level [...] The model today runs every night. Based on what the model predicts, some kind of action is then taken." (Interview C6-3, Pos. 167–177)

Explained in part by the organizational separation or firewall between editorial and publishing in <2015, a parallel data history unfolded within the digital department. Initially, data work was located under the umbrella of audience analytics with <15 people, where the team worked on data infrastructure, data collection and analyses on request. Then, the data science department was added: "As a company, we then realized that we first needed greater differentiation due to the mass of people, but also due to changing needs." (Interview C6-1, Pos. 20–25)

The erosion of the firewall has been accelerated by the realities of professionalized data work, with data teams now increasingly cooperating across media divisions:

> "When I started at the publishing house, my colleague in the print data department had his data spread across millions of Excel spreadsheets. Our bond was not strong enough that I ever saw his spreadsheets. He was doing a good job and was responsible for a lot of subscriptions in that space, but logically, the technical possibilities and the way things are developing not only in-house but also externally are leading to a higher degree of professionalization." (Interview C6-1, Pos. 334–343)

How was data work institutionalized at the publisher? Management recognized the potential to automate recurring data analysis tasks that were previously executed manually by different people. This led to an undesired heterogeneity in the results: "Everytime you do things manually and have different people do it, the outcome looks a little different. And figures and data are no longer comparable. To counter this, the decision was made to establish a centralized data authority." (Interview C6-2, Pos. 466–471) Additionally, with digital subscriptions emerging as a new business objective, the need to gather and scrutinize data became increasingly apparent:

> "Selling a subscription model well and developing it further certainly doesn't work well without data-informed decisions. And this has simply created a business need to invest much more heavily in the area of data and to professionalize around data." (Interview C6-2, Pos. 434–436)

Overall, interest in data and demand for reports have surged dramatically in recent years. (Interview C6-2, Pos. 87) Heightened managerial attention is also reflected in a structural change made in [before 2020]—the editorial head of data is no longer situated within the audience development team but now reports directly to the digital news managing director: "Historically, it's often the case that as head, you usually report to a director who is still somewhere in between. We have done away with that [before 2020]. And as a result, data has now become a central area." (Interview C6-2, Pos. 88–92)

In recent years, the publisher began to focus on churn prediction, statistically examining the factors that contribute to subscription cancellations, the head of subscriptions says. Implementation of these plans was accelerated by the so-called "corona-bump", a significant surge in subscriptions prompted by the pandemic, which eventually stabilized but led to a higher plateau of subscriptions compared to pre-pandemic times. The two data teams collaborate on customer retention across media channels. For instance, the print team identifies an increased statistical probability of a specific customer cancelling their subscription, prompting the digital team to deploy targeted countermeasures on the web to retain them (Interview C6-3, Pos. 366–369). During that same timeframe, there was a noticeable development in the understanding of how their work contributes to digital products: "As far as I can see, I would say that a cultural change is definitely underway. [...] When I started [before 2020], the [digital subscription product] perspective was still very, very weak in the newsroom." (Interview C6-2, Pos. 245–246)

Another structural change was brought about with the hiring of trained data scientists to complement the existing self-taught personnel. These new hires were provided with the resources they required to effectively fulfill their roles:

> "We always had analysts who taught themselves with a lot of enthusiasm and were then able to use the tools. Whereas now we have mathematicians and data engineers on board. That's a huge change and a structural challenge. People work according to the skills they bring to the table. You also have to provide for them technically. You have to be able to deliver clean data." (Interview C6-4, Pos. 385–392)

On a macroeconomic level, the increasing demand for data workers across industries led to a larger pool of specialized talent for the publisher to tap into. Where in the past, autodidacts had "slipped into data", today there are dedicated training programs and applications for data roles are in the hundreds at the publisher (Interview C6-1, Pos. 265–267): "In these moments, you realize that something is changing in general. [In the area of] data merging alone, today you have someone who just goes ahead and quickly assesses the validity of the data. A completely different ball game." (Interview C6-1, Pos. 268–288)

The editorial head of data expresses skepticism toward the broad concept of data science, emphasizing instead the necessity for these technical professions and sub-professions to operate within a specific configuration to make sense at all: "We jumped head first into data science [before 2020] and didn't realize we also had to do groundwork in the field of data engineering. If you make the decision to build up the data area, you really have to think about where to start." Augmenting solid data engineers and data analysts closely in a team appears crucial to the expert, mainly because "data scientists come at such an incredibly high price" (Interview C6-2, Pos. 582–593). Recognizing these growing pains, the head of subscriptions describes the challenges involved in incorporating the roles required for data as "extremely demanding", given the continuous evolution of data work:

> "You first have to understand that colleagues who operate tools are expected to fulfil very different requirements than colleagues who do data engineering and set up scoring models. The demands in recent years are enormous. I mean, the publishing industry might perhaps be a little slower than other industries anyway. It's an enormous challenge to handle this, to recognize the skills you actually need." (Interview C6-4, Pos. 514–522)

But could all the data work done at the publisher be thought of as fundamentally different compared to previous offline marketing efforts? The head of subscriptions affirms the notion that all the data work carried out at the publisher could be considered fundamentally different from previous offline marketing efforts, while her colleagues remain neutral on the matter:

> "In order to do [paywall adjustments] well and to give editorial content its ideal visibility, you need data and that is completely new. The work and handling of data, the derivation from it and the consequences of how content is then played out—that has taken on a new quality." (Interview C6-4, Pos. 569–574)

Looking into the future, the multiplication of distribution channels, social networks and products will only continue, emphasizing the need for further automation and data-informed decision-making (Interview C6-4, Pos. 625–629). Supporting these estimations, the group head of data specifies how demand for data work is anticipated to multiply year-over-year: "We are welcome sparring partners. I would say that the number of data requests has grown by a factor of two to three in the last year alone." (Interview C6-3, Pos. 585–594) As to the specifics of data work, the editorial head of data expects workflows to remain stable while the underlying technology goes through

constant change (Interview C6-2, Pos. 731–737). Somewhat inconsistent, she goes on to predict that the role of the data scientist will disappear completely:

> "I believe data scientists will become less important because more work will be automated. We will probably need less brainpower specifically in these areas and will have to up our data processing game instead. This is at odds with what many other people think. Data science appears to be the absolute non-plus-ultra. But calculating models and such are things that can soon be completely outsourced technically." (Interview C6-2, Pos. 732–769)

Outsourcing of data science would mean drastically changed imaginations around data and reducing data work at publishers from the complex statistics and machine learning models of the data scientist down to solely aggregating and preparing raw data as done by data engineers would relinquish data autonomy back to external actors. Why would they give up autonomy? Cost. External providers "can build scoring models within hours" and "data scientists will be surprised to see how quickly they can be rationalized away, to be honest", the editorial head of data foreshadows. (Interview C6-2, Pos. 769–72) To her, the current state of data work is transitional: "We do everything manually now and I think it's great because we get to know our company really well. But I don't think we'll need it any more in the long term." (Interview C6-2, Pos. 772–776) Echoing this sentiment, the audience analyst expects a quicker transition from manual data analysis to data science: "I believe that manual analysis will increasingly be replaced by data science, meaning that data analysts will perhaps often do the preliminary work and then it moves to data science more quickly." (Interview C6-1, Pos. 395–414) Another change in the quality of data work, data should be less looked at by humans and more worked on by machines:

> "Data will be used less and less for pure observation and will work more and more for us. There's already a development at other companies that are personalizing their websites more, for example. And they no longer do this manually but are building automated systems that react to certain data." (Interview C6-1, Pos. 400–414)

Ongoing data work

Similar to other cases, the publisher tries to optimize earnings by not only attracting new subscribers but also mitigating attrition. Using statistical learning methods, the publisher currently tries to identify predictors for customer churn after customers have successfully paid their first monthly fee:

> "I'd say the most relevant thing at the moment is churn prevention. [...] Over the last two years, we focused a lot on the order funnel and the first 30 days of the subscription. We have collected all the low-hanging fruits there and are now switching to the phase after first payment, which brings its different challenges because customer data is structured differently. That's the focus right now, to develop a model. We are currently in the exploration phase and are finding out how people behave after paying the first bill, [...] to predict whether someone is at risk of cancellation. This is one of the biggest and most complex projects we have at the moment." (Interview C6-4, Pos. 43–61)

Inside the audience analytics team, the introduction of a new tool again underscores the consequences of data work professionalization. With increased capacity resulting from the delegation of infrastructure tasks to the editorial data team, the analytics team is able to develop more technically sophisticated solutions in-house. While certain details remain obscure, these software tools

are said to improve the validation of data hypotheses, as the software would not only crawl the page and extract technical information, but also triangulare with logfiles[146] and information about what Google's crawlers requested—making this combination of data "incredibly attractive in order to be able to draw better conclusions and verify theories" (Interview C6-1, Pos. 165–171).

External factors

When asked about the competitive landscape in publishing, the group head of data characterizes the publisher as "punching above its weight" (Interview C6-3, Pos. 31) in terms of data work. Overall, he sees medium-sized publishers and associations of smaller imprints as less advanced in their approach to data: "What I realized is that there is a difference. And this is not about self-congratulation. [...] We play [at the highest level]. We are already in a different league to many daily newspapers or associations of daily newspapers." (Interview C6-3, Pos. 32–38) He also identifies several organizational factors that distinguish data work at this publisher in a) the level of automation, b) the adequacy of technical infrastructure, c) budget, and d) the redundancy of roles:

> "Cloud infrastructure costs are not the problem, but the operation, the DevOps[147] of these systems, constantly adapting them to when a new service is added, perhaps the loyalty program. We are in a very dynamic environment. [...] What we do is far removed from what I see in small publishers in the daily press. If we look at our marginal costs, they are very low. If we only had a tenth of the subscribers, it would work, but the technical solutions would be too large then." (Interview C6-3, Pos. 229–248)

[146] See also 8.1, "Logging/Logfiles"
[147] See also 8.1, "DevOps"

As a cautionary tale, he describes how a lack of redundancy in technical roles and isolated knowledge can pose significant challenges for smaller publishers:

> "At some point, a lot of time and money will go into automating processes. Whether it's deletion requests under GDPR that are automated. I see a lot in small publishing houses, where there are extremely valuable employees and they are extremely competent and when they leave it's a disaster, because they have done everything and there's no redundancy. That's a level of maturity that we simply have to allow ourselves." (Interview C6-3, Pos. 257–262)

When asked about particular data challenges in the future, the head of subscriptions highlights the growing complexity of data privacy frameworks that need to be navigated and which are demanding for a medium-sized publishing house: "It was already quite demanding with the introduction of GDPR over the last two years. A lot had to be done. So that entailed many, many tasks for us, like asking for complete consent before you even enter the site." (Interview C6-4, Pos. 618–623)

Data and the newsroom

A new and "very exciting" (Interview C6-4, Pos. 267) function, called the subscription editor, interfaces between the subscription and editorial teams, and directly makes data-informed, editorial decisions on certain elements of the homepage and other pages as things happen, a task that "entirely depends on data". (Interview C6-4, Pos. 307) This editor would control a box on the homepage displaying teasers to subscriber-only articles from the archive which still generate sales or conversion reliably or become relevant again—a method

called “resurfacing”, which now accounts for a relevant proportion of sales and is data-based.” (Interview C6-4, Pos. 307–315)

The necessary tools and interfaces to identify content suitable for resurfacing were subsequently built by the editorial data team:

> “[We are building] tools to monitor subscription content, tools to identify suitable content that could be put behind a barrier.” (Interview C6-2, Pos. 258–260) The subscription editor typically operates with data at a deeper level than other editors: “There are people looking exclusively at the top level and at the results that are presented there. [For example], an editor or managing editor and the department heads. On the other hand, of course, there are also people who work at an imperative, deeper level, such as the subscription editor, where they can perhaps create dashboards themselves and delve deeper into analyses.” (Interview C6-2, Pos. 322–331)

Here, the subscription editor becomes a boundary worker in the clearest sense, not only receiving and acting on data in a curatorial way but also interacting with the data affordances made available to him by the data team. On the question of unit boundaries, we see data work interpreted as an equalizing force. Overall, there exists a shared responsibility towards a digital product. Data affordances are essential for editorial observation, and editorial observation is crucial for product success—a situation which fosters a sense of co-dependency between the product, editorial, and data teams. A new degree of collaboration across boundaries appears to be a recent development and the “constant dialogue between data, paid content, editorial and editor-in-chief” is understood as a “sign that there is a paradigm shift or close cooperation and a shared understanding of where we want to go” (Interview C6-2, Pos. 263–266).

With many emerging data-oriented roles now integrated into editorial teams, the separation between publishing and editorial has mostly disappeared, the head of subscriptions asserts: "In my opinion, the editorial team and the publishing house have completely dissolved. It no longer exists in this form. There are many, many more roles that have a focus, on one side and on the other." (Interview C6-4, Pos. 414-417) Overall, there has been a notable rise in interest in data, to the extent that regular updates on performance metrics have become the norm. However, such a deep immersion with data remains voluntary as the editorial team has become more data-informed and anyone can look into data: "Everyone has a transparent insight into metrics and data if they wish. It is available to everyone and there is a review every fortnight, for example, in which all the figures are disclosed." (Interview C6-4, Pos. 418–422) With this data transparency policy in place, data cannot be described as opaque, nor is access to data restricted (as in other cases), even though the data workers acknowledge that with they now generate an "incredible amount of data that some people are just not interested in." (Interview C6-4, Pos. 442–446) Overall, the publisher seems to adopt a rather egalitarian approach to data and data access as reflected in a daily data report sent out to the whole staff. (Interview C6-4, Pos. 447–451)

Decision-making with data

On the question of data-informedness versus data-driven decisions, the head of subscriptions says how "everything we do in journalism is data-informed" (Interview C6-4, Pos. 330–331). As they provide their colleagues with "as much information as possible" about articles and department heads have reference values they can translate into good or bad days, there are no target quantities any department: "That in turn would be [data-driven] for me." (Interview C6-4, Pos. 332–337) Conversely, within his own domain and interfacing with

internal customers, he defines the mode of operation as "truly data-based" or data-driven: "If we say we want to improve the ordering process, then it has to be proven with numbers. But in journalism this is borderline unthinkable." (Interview C6-4, Pos. 337–340) Overall, data has provided measurability and explainability to recent successes in digital subscriptions. In this way, successful decision-making seems to be both created by and illuminated by data:

> "Work in the area of data has made it possible for product management to make decisions and improvements that are clearly measurable. [Management] realized that in order to be able to measure things even better and make even more granular improvements, you need an even deeper understanding of data." (Interview C6-2, Pos. 562–566)

As part of the regular routine at the publisher, hypotheses about incremental product alterations are tested through statistical experiments and control groups. Anyone within the organization can contribute hypotheses and, apparently, these hypotheses are gathered and catalogued by the paid content team. Improvements in this area appear to correlate with improved revenue or metrics directly associated with revenue:

> "Last year, we ran around forty to fifty A/B tests on the purchase funnel. The first one hundred days or the first thirty days of the subscription are super important to us. [...] For example, we have tested whether to send out daily or weekly reading recommendations for [subscriber content]. This is then tested against engagement. And the result was that a weekly reading recommendation with five articles have a stronger effect than a daily recommendation with just one article." (Interview C6-4, Pos. 129–168)

With entirely new products, on the other hand, A/B tests are less important and these are built and launched on the basis of creative intuition and survey data: "As such a product is conceptualized and designed, there's a lot of newly created code. Naturally, not everything is tested beforehand, but there might be a reader survey in advance or a user shadowing with customers." (Interview C6-4, Pos. 112–127)

As often mentioned by interviewees, reports in the form of non-interactive digital data artifacts circulate at the publisher as a basis for decision making. These typically include an interpretative layer, presenting not only plain metrics data, but also data filtered by key insights and expressed as natural language inside a visually appealing presentation:

> "There are wonderful reports for management and other stakeholders. These are nicely presented in a graphical way. We try to be very descriptive so that you can quickly recognize how the month went. For example, what's the ratio of engagement to paid subscribers? What's the absolute number of cancellations in the previous month?" (Interview C6-4, Pos. 277–283)

Aside from these scheduled reports, the editorial data team tries to build automated systems that act on data as it emerges, a mechanism also called event-driven[148] data processing. The group data team, on the other hand, operates in 24-hour cycles as they "dump new information into [the systems] once a day and retrieve responses nightly" (Interview C6-3, Pos. 346) which also constitutes a difference to the web context, where "of course the moment the customer logs into their user profile, [you] have to know who they are, and deliver the right message to them" (Interview C6-3, Pos. 347–361).

[148] See also 8.1, "Event Analytics" and "Event Pipeline"

The timeframe of data cycles here becomes an important inflection point, where the "real-time" display of data requires "real-time" agency:

> "The difference is that we cannot intervene directly on data. We just send out a mailing and things happen after the fact. [In our case] presenting data in real time doesn't help you because it's only ever imaginary real time anyway." (Interview C6-3, Pos. 398–409)

Metrics and data sources

In general, data and metrics play a "hugely significant" (Interview C6-4, Pos. 85–86) role in the daily routines of the people working on the paid content team: "[We] have insights into the product, a complete digital product. So, we can measure and see a lot. And of course, we also use this to constantly improve and make things more user-friendly, expand and extend them." (Interview C6-4, Pos. 85–100) Differentiating metrics into two subsets, the editorial head of data views key performance indicators as those metrics utilized for controlling and steering a company: "Standard or general key figures are mostly relevant at the granular level. In the places where the data on which these figures are based on are actually collected." (Interview C6-2, Pos. 195–200) In addition to the more obvious quantitative metrics, the paid content team also generates qualitative data through surveys: "Another layer, of course, would be insights provided by surveys, meaning that we can also discover things using qualitative methods or standardized surveys, things we can't measure with tracking data." (Interview C6-4, Pos. 85–100)

With the concept of data products, internal applications that rely fundamentally on data modelling and provide interactive access to specialized metrics, staff receive insights into "standardized data" around customer

behavior (Interview C6-4, Pos. 536–541). As an example, by measuring the frequency and recency of activity among new subscribers, the publisher is able to model for the dependent variable of *likelihood to pay*. Using statistical prediction in this way, the publisher bridges a gap in data availability:

> "A scoring runs across all technical formats, with each user collecting points every day. These are based on frequency recency. When was the user last there? How intensively do they use, how deeply do they use? How long do they read? Then there are bonus points for sharing, commenting and the like. And this score, totaled each day, correlates significantly with the dependent variable of payment after four weeks. Which makes it a great tool for us in product development, because we don't have to wait very long to see if the beta group pays. [Normally] it takes four weeks plus a delay due to invoicing and billing." (Interview C6-4, Pos. 144–160)

Delving further into the concept of user engagement (the level of use and signals of interaction left by users), the head of subscriptions considers the metric as a useful "buzzword" to encapsulate complexity. In the context of a digital application, she explains, the publisher tracks success across other significant factors to somehow measure activation, and "while you could also call it retention", engagement remains "the most important KPI of all" as it becomes "obvious that someone who uses their subscription is more likely to stay" (Interview C6-4, Pos. 183–188). Other important metrics mentioned by the head of subscriptions include conversion rate, cancellation rates, and open rates as well as click-through rates for newsletters (Interview C6-4, Pos. 189–195).

The audience analyst reiterates the evolution of influential metrics, which, at this publisher followed a path similar to that of comparable digital news organizations: “In the past, like our competitors, we used to look purely at sessions and historically at impressions. [...] Currently, the most important metric in generating reach is the number of engagements per content [Einstiege].” (Interview C6-1, Pos. 85–97) Starting with the key performance indicators of visits and sessions, she then inspects these metrics across their sources of “acquisition”, referring to the place users came from immediately prior to navigating to the publisher’s pages and articles (Interview C6-1, Pos. 194–199). Cross-referencing sources of acquisition and the categories of content would provide insightful data. However, to establish this kind of complex metric known as *performance per channel*, more groundwork would be needed (Interview C6-1, Pos. 200–204). The underlying raw data points are called tracking data here, implicit data from user activity, which remains “the most important data source” as the “movement data of our users is super relevant” (Interview C6-2, Pos. 127–135). Explicit data, which refers to data intentionally entered by users, along with other types of data, are regarded as secondary sources currently not meeting the standards, but the data workers are nowhere near as far as they would like to be in terms of such secondary data: “It’s not like we had super good data streams already” (Interview C6-4, Pos. 224–225). Lastly, the publisher aims to digitize explicit data collected during phone calls or other human interactions by “agents”, first-level customer service employees, to then further automate processes: “Perhaps [the customer] will cancel our service because they haven’t received their latest issue due to snow and ice. A signal that could be telling us: watch out, cancellation is imminent.” (Interview C6-4, Pos. 235–241)

Digital and physical metrics are distinctly separate concerns, as the group head of data confirms. His unit focuses on generating and monitoring print subscription metrics, including overall circulation, quarterly circulation, cancellations, and implicit cancellations due to contract expirations. Another set of metrics allows the group-level data team to track the effects of marketing efforts, which subsequently impacts the number of print subscriptions. These metrics primarily revolve around the number of "responses" and "orders" (Interview C6-3, Pos. 277–288). Just like its editorial counterpart, the group-level data team monitors email performance metrics like open rate and open time:

> "[We gather] a lot of transactional data from emails. So, after we have delivered a marketing campaign, did you open the email? When? These are data points which allow us to segment accordingly, so this email channel is very, very relevant for us." (Interview C6-3, Pos. 326–331)

Once a print subscription has begun, billing and fulfillment are outsourced to a service provider. This service provider then hands over related customer and billing data through an API. This data source constitutes the bulk of the overall group-level data collection. (Interview C6-3, Pos. 331–338) All data and metrics discussed above fall under the broad categories of systems data and user or customer data. Additionally, the publisher possesses article metadata for organizing articles into sections and making them searchable. However, identifying the success factors of successful articles has proven challenging (as in other cases discussed in this study). As a result, this endeavor is currently on hold:

> "We tried to unleash standardized success measurements on our data. Can we find the golden formula? What makes an article successful? What are the characteristics? We didn't succeed. At least not with what we have at our disposal. Accordingly, we have put the analysis of article features on the back burner, [...] we are concentrating on user behavior instead. However, I believe that there is something to be gained there. You just have to tag article metrics differently or make them measurable in a different way or record them in a standardized way in order to then perhaps delve deeper. What influence does the tonality of articles or the design of images have? Or the wording of headlines? We still work with a well-developed gut feeling here." (Interview C6-4, Pos. 636–655)

Overall, as observed in other cases, the publisher has an overabundance of data points and one major challenge lies in making sense of all this data, as exemplified by a catalogue for the cancellation phase with fifty individual data points: "But what's actually in there? What is truly measurable? And then we look at which metrics have a better explainability. We can measure a lot of things but the challenge is to find out whether this helps us to make progress." (Interview C6-4, Pos. 195–202)

Dashboards

A major use case of dashboards at this publisher involves displaying systems metrics or "health" data. Interestingly, the head of subscriptions mentions these systems dashboards first, implying her sense of ownership for technical questions as well (Interview C6-4, Pos. 254–260). Some of these dashboards have been integrated as bots into the company-wide communication platform Slack. This might indicate how dashboards, data artifacts with a relatively

clear shared meaning but limited immediacy, will in the future give way to more direct means of communicating data, like corporate chat applications:

> "I no longer look at the alerts because we have integrated them into Slack. In other words, if there are outages and certain KPIs are broken, such as [the HTTP status codes] 404 or 403[149], certain pages are not deliverable, not accessible, then Grafana pushes us via Slack and alerts and de-alerts quickly." (Interview C6-4, Pos. 269–277) The editorial data team provides visually pleasing reports to business stakeholders, operates, and updates conventional spreadsheets, but dashboards are not a crucial item to them (Interview C6-2, Pos. 337).

On the publisher level, the concepts of "dashboards" and "reports" overlap as both are means of reporting in a managerial sense. The team's work is split across four categories: 1) automated performance reports via team chat 2) longer automated reports via e-mail 3) self-service reporting dashboards, and 4) ad-hoc analysis requests, with reports taking up more effort than the dashboards:

> "We have now switched to commenting data and are doing quite a lot of work. For one thing, the group of recipients is very large for these reports. And it's no secret, this group is not necessarily the most analytical. [It has to be a] well-formulated, formatted report with graphics, everything nice. We provide a handout here." (Interview C6-3, Pos. 437–485)

According to the audience analyst, the distinction between dashboards and reports, whether manually created as one-offs or through "individual analysis",

[149] See also 8.1, "HTTP Status Codes"

or analysis repeated in intervals and semi-automated, becomes clearer. Surprisingly, the demand for individual analysis remains much higher compared to automated reporting: "You don't have to constantly look at everything all the time, but things are instead more interesting on a project-by-project basis." (Interview C6-1, Pos. 236–238)

Overall, having a structured set of data to manually query for answers and report on seems more critical to daily business than the automation of data processes. Such ad-hoc analyses are "mostly run only once" and so "it's not worth automating", but a case where one "simply needs a very solid database" (Interview C6-3, Pos. 524–525). Queries mentioned: How long does it take from expression of intent to cancellation and then until a cancellation is actually processed? Do customers acquired during the pandemic cancel more frequently than others? (Interview C6-3, Pos. 524–528) Dashboards are a less talked about concept compared to other cases, with manually created reports taking precedence over dashboards, while the generally enthusiastic sentiment towards dashboards in the field is acknowledged, but with a degree of distance: "In dialogue with colleagues from other companies, I've noticed that dashboards are really en vogue right now." (Interview C6-1, Pos. 495–502)

Technology and tools

On the storage level, the publisher uses BigQuery[150] as the central database where all disparate data sources are stored and all kinds of queries are executed. Although analytics software Webtrekk[151] still generates and holds a significant portion of highly relevant tracking data, a migration towards the Google Cloud Platform and BigQuery is currently underway:

[150] See also 8.1, "BigQuery"

[151] Now part of Munich-based company Mapp.

"We want to mitigate the situation because Webtrekk should only be a data source and not a data management tool." (Interview C6-2, Pos. 168–178) The editorial data team uses two user-facing software products to build reports and dashboards, Webtrekk and Data Studio (Interview C6-4, Pos. 253) Systems dashboards and alerts are built with the open-source dashboarding utility Grafana (Interview C6-4, Pos. 268). The editorial data team also works with conventional spreadsheets, which the editorial head of data considers to be legacy technology—meaning outdated, but still operational—as there are still reports that were historically created in Excel or Google spreadsheets (Interview C6-2, Pos. 339–340).

A growing professionalization of data work, characterized by the move of data concern from audience analytics into multiple departments, is evident in the evolution of tooling. Previously, data storage and presentation were facilitated with editorial analytics software whereas now various sources are increasingly visualized with Data Studio and pooled in a proprietary data lake:

> "Here I'd say a major movement is currently underway. I think that in [the future] we will generally be looking at a lot of things in a Tableau or Data Studio because we increasingly need to bring together data from different systems." (Interview C6-1, Pos. 157–161)

On the publisher side, the software set reflects an emphasis on displaying data to stakeholders, with two main external services aimed at large corporate marketing divisions in a) Emarsys, a "customer engagement platform" owned by SAP, and b) Episerver, an email automation application. (Interview C6-3, Pos. 346–352)

Notably, the group-level data team uses different software than its editorial counterpart in their visual reporting for marketing and sales colleagues with QlikView, a "vast universe where you can display [figures] interactively" (Interview C6-3, Pos. 491–499). In addition, the editorial head of data highlights Linkpulse as the example of third-party software that significantly influences data work in general, where dashboards are consistently used in day-to-day editorial operations: "[Linkpulse] is not meant for reporting, but rather for monitoring. They provide real-time dashboards about what's happening on our website. Which articles are performing well? Always in relation to how much reach or subscriptions they generate." (Interview C6-2, Pos. 277–278) A less time-critical task involves the team combining past and present article performance data in a new tool to identify which older articles they might want to resurface: "[We are developing] tools to search within our inventory, old articles, in order to identify so-called evergreens, for example. Articles that have performed well again and again over time." (Interview C6-2, Pos. 284–286)

6.3 In light of theory

6.3.1 Exploring the evolution of data work

Building on the case analysis subtopic of "origins and changes", we can envision the individual timelines of data work and compare between cases. This directly ties in with the first part of RQ1.

Research Question RQ1

How have news organizations shifted or enhanced their ways of working with data in recent years and can we identify structural or technological patterns on an inter-organizational level?

I want to concentrate on the chronology of changes and postpone technical details for the moment. At case study 1 (C1), the publisher's shift in thinking about data and metrics can be traced through three historical phases: measuring advertising revenue based on reach, experimenting with composite metrics, and adopting key metrics that align with new operational goals around digital subscriptions. Structural changes have led to the emergence of new units and roles such as data scientists and data engineers, referred to internally as "data persons" (Interview C1-2, Pos. 158–159). Driven by economic pressures and the influence of data narratives originating from Silicon Valley, the concept of data science as a work profile was introduced relatively recently. Initial awareness and utilization gradually grew within the organization over time. However, prediction and machine learning models are not yet fully integrated into production. With the establishment of centralized data competence, some data work has also shifted from editorial analysts, who traditionally embodied advocacy for editorial interests and helped with web analytics, to the more technical oriented data analysts (who are no longer part of the editorial team).

Another, albeit less significant change, is the availability of raw data, which has enabled the development of what the organization refers to as "data products"—the specific properties of which remain unclear. All of these changes have occurred within of the last <5 years.

At C2, the shift in organizational focus from traffic to subscribers is portrayed as a response to the volatility of search engine traffic, which is referred to as "sweet poison" due to its unpredictability (C2-2, Pos. 307). This shift is seen as a way to regain autonomy and independence for publishers. In turn, this shift, along with external technological advancements, such as ecommerce and increased trust in online transactions, as well as internal cultural (measuring customers, not articles) and technological changes (adjusting the paywall required proprietary data), led to the establishment of a centralized data department (which includes various data-related roles down to some financial control) and a focus on customer-centric approaches. Within this newly established data department, purely operational data work like financial controlling is conducted alongside editorial analytics, reporting, and general tasks related to dashboarding. A change in culture and the technological conditions within and around the publisher (such as the acceptance and popularity of digital payments) paved the way before any strategic or deliberate change management towards the professionalization of data work took place. Both indirect and active management of such change is attributed to the executive leadership, from board members to marketing leaders. Additionally, teams began to self-regulate based on Objectives and Key Results (OKR)[152] around 2020, before the methodology was formally adopted by the whole organization recently. Again, these changes have taken place within the last <5 years.

[152] See also 8.1, "Objectives and Key Results (OKR)"

Case 3 (C3) adopted a membership-based business model without collecting data, but later recognized the need for improved data management practices. Having acquired analytics software without fully understanding its functionality, this led to challenges in data attribution (to user profiles) and interpretation. Over time, the startup invested in a database/spreadsheet tool, faced challenges in data harmonization and normalization, and eventually limited their scope to authenticated users to reduce data to more manageable quantities. The initial wide-scale data collection here began <5 years ago.

After initially starting their data efforts with the goal of consolidating units and tapping into pre-existing data resources, management at Case 4 (C4) observed in <2020 that the national publishing industry was relatively inactive in terms of data efforts. This prompted them to acquire consultants and explore established data strategies internationally. An international transformation program thus played a significant role in driving innovation and shifting the focus towards measuring digital business performance. In <2020, a statistician was hired as the CEO, and the advisory board was replaced with product and data experts. All company-wide data practices then involved several phases and spanned 3 years. Initially, the decision to pursue digital subscriptions led to the promotion of a head of data who began exploring available data sources, prospecting data infrastructure, and building a team. One year was dedicated to assessing, planning, and cleaning the data, while simultaneously building a data lake to harmonize and manage the data. The next year focused on expanding the data infrastructure and delivering results, which included the development of editorial dashboards. In the year leading up to the interviews, experimentation with machine learning, particularly in natural language recognition and prediction, was carried out. Overall, data work became ingrained in upper management, with a belief in the importance and

indispensability of data. New roles emerged after >2020, notably the business analyst and the data analyst. Similar to other cases, all steps, from observation, acquiring data consulting, founding of a centralized data company, and adapting the organizational framework of OKR all occurred within the last <5 years.

At Case 5 (C5), dedicated data workers were initially integrated within the editorial team but found to be unproductive. Around 2020, a second attempt was made by the head of product, resulting in the establishment of a team focused on data analysis and research. In order to appease a strong editorial coalition, they positioned themselves as qualitative researchers, providing answers to complex questions based on data. Operating within an organization dominated by editorial, several steps had to be taken to legitimize the new approach. These included lobbying at the highest level and presenting their work in a more narrative way. However, despite these efforts, the team still faces skepticism regarding the impact of data and doubts about its potential for continued growth. One interviewee hints at a near-term change, with the team potentially shifting their reporting from product to sales—a step that would distance the team even further from editorial, at least structurally. As the fourth case in the sample, the organization has also adapted OKR in recent years.

Initially, data work at Case 6 (C6) started with web analytics aimed at enhancing digital product reach and advertising sales. The introduction of search-engine technicians with expertise in web analytics marked the beginning of data-related practices. The pressure to deliver data-informed insights increased due to the business model of digital subscriptions and its direct impact on revenue. Initiated by upper management's observation of

inconsistent form and logic in data reports as early as <2018, the first steps towards in-house data infrastructure were taken with a CRM project. This initiative eventually led to the establishment of a data team and a data warehouse. At first, the separation between editorial and publishing departments hindered data collaboration, but since then, professionalization and cooperation divisions have increased. With the increasing relevance of data work, a structural change elevated the head of data to now report directly to the digital executive. According to interviewee accounts, editorial staff have since developed a product-centric mindset, recognizing the impact of data on digital products. Structural changes were implemented by hiring trained data scientists to complement self-taught personnel. Alongside this professionalization, data practices were found to be sequential processes, with data engineering and data analysis roles required by and preceding the work of data scientists. While interviewees expect data workflows to remain stable, the underlying technology might eventually move outside of the organization again, ending the current in-house data science efforts and a greater emphasis on data groundwork at the publisher.

Overall, with C6 as a slightly earlier outlier, the survey confirms how fundamental changes in data practices, organizational structure, and management culture have played out over the last few years. In five cases, new roles were introduced across the data spectrum, with a minimum of one data scientist and additional data engineers, data analysts, or business analysts in multiple cases. Except for the startup, all organizations in the sample have established a dedicated data department over the last five years—even a whole new company dedicated to data work at C4. In each case, changing data practices can be traced back to a re-orientation towards (digital) subscriptions or customers.

I find multiple matching narratives of increased attention to and investment in data as a result of market forces devaluing traffic or reach, pushing the organizations to embrace a business model based on digital subscriptions. In four cases, the Objectives and Key Results (OKR) management framework was adapted around the same time as the structural changes towards greater centralization of data work occurred. In all cases, data work that historically happened elsewhere in these organizations (e.g. financial controlling, search engine optimization, or editorial analytics) moved into centralized data units to some degree.

Addressing RQ3, I aim to explore how the qualities of software-enabled data work may differ or be inherently comparable to what media organizations have done before the advent of the internet:

Research Question RQ3

How does data work in its current form differ from previous ways of working with data, given the assumption that data work does not require digital affordances per se?

At C1, interviewees question whether data work can be regarded as more consequential than it was pre-internet. While data is increasingly valued, there is a risk of prioritizing quantifiable objectives over other important factors, potentially even stifling innovation. Interviewees at C2 were divided on the matter, with some viewing their data work as a return to traditional marketing with the aid of current technological means. Others believe that these exact technological means available now are so vastly different that the quality of data work should also be considered as vastly new as well.

Positions at C3 highlight the unique nature of data work in the digital era and express skepticism about the affordability of advanced data software and data scientists for journalistic organizations. Going beyond the matter of newness, one voice here questions the *raison d'être* of complex data software in the field. The head of data at C4 views their work as a continuation of offline data practices but on a larger scale. At C5, positions are uniformly moderate, data work with interactive media seen as a continuation (since the television industry has long tracked audience behavior) or an evolution (due to improved tools and larger datasets), but similar underlying questions.

6.3.2 Encountering data assemblages

Next, I will deconstruct and compare the various data assemblages encountered in the sample. As established in Chapter 2.2.1, these data assemblages are theorized to consist of two main components—a technical stack and a contextual stack. Starting with the technical stack(s), I will critically examine statements about infrastructure, software, and proprietary data affordances as collected in the case analysis sections under "technology and tools". This aligns with research question RQ2 stated in Chapter 4.2:

Research Question RQ2

How are data generated, processed, stored, shared, and analyzed and what can we learn about the specific infrastructure, software and data affordances used to facilitate these activities?

At C1, dashboards are a pervasive concept, with the term used regularly in all interviews. On the other hand, reports as a device for sharing data are discussed less. Dashboards are attributed with a real-time quality here, thought of as having immediacy, accuracy, and relevance. Such a requirement from editorial might explain how dashboards are primarily used for traffic management. Another term that carries a similar sense of immediacy, dashboards are said to receive "streaming" data from the paywall software CeleraOne and are displayed through a service called Chartio.[153] Other technical concepts discussed extensively are prediction with ML, though this line of analysis remains relatively inconsequential for the moment. The team primarily uses Big Query, along with open-source tools such as Python and Scala.

[153] A testament to the pace of technical evolution, both software products have since been bought (by Piano and Atlassian respectively) and integrated into other data software.

Still actively used by the editorial team, the analytics tool Linkpulse is not mentioned by the data team, suggesting a shift in focus towards custom reporting and dashboards for sales and marketing.

At C2, data affordances are discussed in comparison to legacy systems like SAP, with the new tooling seen as delivering faster access to data, more up-to-date data, and automation of data flowing between systems. Representing a mindset around data primarily as a tool for financial controlling or operational steering (and editorial analytics as a nuisance), here the prevalent notion is data as an enabler of so-called business intelligence (BI). Consequently, dashboards are seen as a desirable means of automating away the need for human interaction with the data department—a unique assessment in my sample. In turn, new data tools have facilitated the construction of a BI infrastructure, as demonstrated and explained in extreme detail by the head of data. In terms of data software, the publisher plans to transition to Adobe Analytics for editorial purposes. However, data visualizations for sales and marketing are realized with a comparatively more complex technical stack based on Microsoft Power BI—indicative of a data department operating with two classes of technical stacks. C3 relies on the *software-as-a-service* (SaaS) database Airtable as its central data repository and interface, facilitating various data work processes. Given C3's small size and limited resources, the choice was motivated by the cost and complexity associated with developing proprietary data solutions. Consolidating data from multiple SaaS sources before integrating them into Airtable, the organization delivers a single dashboarding solution with one unified access level to all its staff. In following a startup mentality that encourages improvisation, the organization faces some undesirable side effects with unmanageable or broken data (Interview C3-1, Pos. 94–95; Interview C3-2, Pos. 318–333).

Interviewees at C4 gave particularly detailed and ostensibly knowledgeable insights into the procurement of their data technology stack, which appears especially sophisticated and is currently spread across all three large cloud providers: Microsoft, Amazon Web Services, and Google. Unique within the sample, C4 also developed a) a custom analytics solution based on the open source software Snowplow, and b) a relationship graph across news articles and topics, incorporating external data by crawling various sources. These are extremely resource and cost-intensive projects, not found elsewhere in the sample and indicative of a heightened commitment to data. In terms of machine learning, interviewees here believe prediction of churn on digital assets to be pointless—a sharp rebuttal to all other positions in the sample.

Tooling at C5 includes Adobe Analytics and Microsoft Power BI for static reporting. In addition to these tools, the data team here stands out in the sample in emphasizing qualitative data work with academic software such as Questback for surveys and Excel Starter or MaxQDA for qualitative analysis. Dashboards play a significantly lesser role here than in the case of other data teams, as they are predominantly used in the editorial department, with no mention of KPI dashboards for sales and marketing. According to the head of data, dashboards provide limited utility and could even be shut off entirely without significantly impacting operations.

C6 utilizes Google BigQuery as the central database for storing and querying disparate data sources. However, a significant portion of tracking data is still generated and held by Webtrekk, which the publisher aims to transition into BigQuery. The editorial data team employs Webtrekk and Google Data Studio for building reports and dashboards, while Grafana is used for systems dashboards and alerts.

Additionally, conventional spreadsheets, Microsoft Excel, Google Spreadsheets, Qlik, Tableau and Microsoft Power BI are used. The professionalization of data work is evident in the shift towards visualizing multi-faceted data from various sources using Google Data Studio and similar tools. On the publisher side, Emarsys and Episerver are employed for customer engagement and email automation, respectively. Linkpulse is highlighted as a third-party platform that provides real-time dashboards for monitoring website performance. The data team also develops custom tools to search and identify older articles with consistent performance.

Overall, a) dashboards are the ubiquitous data affordance worked on, worked with, and talked about, often in conjunction with the underlying notion of immediate and actionable access to data. Report or, interchangeably, reporting, are the next common data artifact mentioned; b) attention or investment in dashboards does not correlate with company size; c) there is ongoing experimentation with data and data infrastructure in all cases; d) machine learning is actively deployed in four cases for churn prediction, in one case exclusively (and counterintuitively) for print subscriptions; e) Google products dominate storage and analysis, especially data-specific products like Data Studio and BigQuery, prevalent across the sample f) in terms of platforms, three cases use Google Cloud Platform, one uses Amazon Web Services, and one uses Microsoft Azure, with various "multi-cloud" configurations; g) there is a notable absence of privacy discussions around data on premise, meaning at the publishers' facilities; h) there are parallel technical stacks in multiple cases, where marketing and sales use more sophisticated tooling than their editorial counterparts.

Next, I attend to the second part of the data assemblage, the "discursive and material components related to philosophy and knowledge, practices, stakeholders, actors" (Kitchin, 2022, p. 23) that constitute the contextual stack. As established in Chapter 5.3.3, I follow along several topics and other types of contexts put forth in literature about data. Somewhat linked, I discuss metrics, key performance indicators and other types of conceptualized data in the sample; how such data and metrics circulate through organizations (Beer, 2016); and trace how organizations establish technological imaginaries (Beer, 2019).

In the case of C1, the overall focus rests on a conversion funnel, reflecting the main managerial goal of increasing the number of paying subscribers. Metrics of interest include impressions by subscribers, impressions by non-subscribers, and conversions. Such ecommerce-inspired metrics are becoming more relevant to data work with interviewees considering them as "standard". Consequently, efforts are made at the corporate level to centralize repeatable data work and transform it into automated services as much as possible, raising concerns about data agency and responsibility at the publication level. These competency claims create a data hierarchy: Once integrated into editorial teams, analyst work is often transferred to the data team, leading to some discontent. Key metrics are determined by sales, marketing, and executive management. In essence, management prioritizes measuring and optimizing for subscribers (with all other key metrics linked to subscriber counts). Furthermore, there is a growing tendency to centralize these functions at the corporate level.

At C2, from a structural standpoint, the establishment of a central department tasked with handling all kinds of data work stands out. Particularly notable is the emphasis on financial controlling, and along with it, a focus on sales metrics like CPO and CPI. All critical metrics appear to carry some relation to the contribution margin—the supreme metric. As the practices of counting, quantifying, observing, reporting data, and advocating for data work coalesce into a unified responsibility, a self-sustaining loop materializes—although not in the narrow sense as defined by Hacking. The data team evolves into a normative force in its own right as users disappear behind their role as potential customers, a function of their qualifiable waypoint as they pass through the marketing funnel. A powerful data imaginary carries both panoramic and prophetic qualities. Panoramic, as interviewees envision a comprehensive "golden record" for every user, akin to the notion of a single source of truth mentioned in other contexts. Prophetic, as the organization aspires to use data to predict subscriber churn. Another interesting practice involves metrics that reveal their true significance only within the context of so-called "shadow quantities", measurements that require further quantification and evaluation based on underlying metrics under supervision by a select group of individuals. With membership data at the center of journalistic production, C3 interacts with its members as domain experts and collaborators, classifying them by geolocation, interests, and other criteria (an interesting avenue for further studies). Evidently, management permits the coexistence of multiple data discourses. On one hand, the small organization lacks dedicated data workers, but still adopts a mindset centered around conversion funnels (a "beautiful metric, because it can be directly expressed in money", Interview C3-1, Pos. 411–412). On the other, interviewees openly acknowledge how their quest for correlations remains experimental and carries with it a risk of over-interpretation.

Overall, interviewees appear divided in their assessments of what data can and can not do—with the senior developer taking the least technocratic point of view. Reflecting a culture of egalitarianism and experimentation, data and metrics circulate freely throughout the organization. Editors and other personnel casually access the same unrestricted data affordances, as data here lacks the discursive power observed in other cases.

At C4, the executive considers the organization's overall digitalization as complete, with the current focus shifting towards leveraging analytics data to enhance hyperlocal and special interest content. This next evolutionary step is driven by the overarching objective of achieving "full local authority over data" (Interview C4-1). In turn, these high standards are also applied to data work, which aims to emancipate itself from merely facilitating paid and premium offerings to encompass "advanced data science"—from building dashboards (which are thought to represent the past), to predictive analytics and data-driven automation. These aspirations culminated in the foundation of a dedicated company for data and analytics, enabling its data workers to concentrate on the critical tasks of analysis and enablement. While the "what" of data questions remains a subject of negotiation with the stakeholders, the data team has authority over the "how".

In terms of management metrics, it is evident that C4 submits all operations under the OKR regime, with the data team offering interactive tools to establish and monitor these OKR metrics. Again, the alignment of advocacy and implementation of data work coincides within a team directly reporting to upper management, catering to, and influencing, all kinds of stakeholders.

As the only example of a wider data ecosystem in the sample, C4 is involved in a data collaboration among regional publishers. Here, a data interoperability format is developed so that anonymized data can be shared to then formulate "joint hypotheses" (Interview C4-1) about data—arguably forming a data standards body (another interesting avenue for further research). This collaboration applies a universal performance quantifier that exerts a normative force across the participating regional publishers. C4 appears to be a notably influential participant in this context, able to establish a form of metric dominance. In terms of metrics, data points are consolidated into a single "North Star" metric of daily active subscribers. The conception of such a comprehensive metric aims to provide editors with a clear focus and prevent them from being overwhelmed by the abundance of data—which otherwise "accumulates so much knowledge, you could go overboard with it" (Interview C4-1). Managing editors engage in weekly introspection sessions, followed by the dissemination of written reflections on how and why their key metrics tilted in either direction. Overall, I find a significant surrender to the data or metrics regime, deeply ingrained in routines across the entire company. Metrics here are used in unique ways, as evidenced by the remarks regarding dashboards as backward-looking, instead serving as gauges of progress towards an imaginary future. Within this framework, the metric contains a planned obsolescence—relevant only until the organization reaches its implicated target quantity. Expanding on Hacking's ideas, data not only possesses prophetic qualities, the specific quantifiers or metrics are overtly performative of the organization's aspirations.

At C5, conflicting narratives emerge around the nature and structure of data work within the organization. On one hand, the organization has centralized its data expertise in a data and research department, which employs a

combination of qualitative and quantitative methods. Yet other key personnel in the organization fail to mention the existence of the department, claiming that the location of data workers "no longer plays a role" (Interview C5-2, Pos. 306–310). Nevertheless, descriptions of hierarchies at data and research are characterized as informal and "customary", suggesting a relatively casual atmosphere. While upper management exercises mild control through quarterly steering panels, the department retains overarching authority over the substantial data-related tasks. Conversion and engagement metrics receive less emphasis compared to other cases. Notably unique in the sample, while the publisher continues to experiment with composite metrics derived from dozens of variables, the data team now advocates for a reading depth metric as a proxy for quantifying the performance of individual articles. The flow of metrics within the company is fragmented, as sales, finance, editorial, and research each operating independently without shared workflows or data dissemination mechanisms. Some departments work with dashboards for their data needs, while others rely solely on static documents exchanged via email. The organizational structure seems to reflect communication and data flows at C5—there are conflicting narratives and a less developed data imaginary than in other cases. While the results might have varied if I interviewed a sales representative, the overall pattern of heterogeneity would likely have persisted. In total, three data teams coexist at C6 with two dedicated units within the publication and one small team reporting to corporate marketing. The head of subscriptions characterizes the organizational style as "structure through non-structure" (Interview C6-4, Pos. 66), although it is technically described as a horizontal or matrix configuration (both terms are used interchangeably). An overarching emphasis on communication and negotiation stands out, as interviewees often speak of "receiving" and "listening" rather than "sending" (of reports, advocacy, education or enablement).

Unique within the sample, this is also reflected through the addition of data service personnel, who serve as mediators between engineering teams and stakeholders (akin to requirements engineering[154]). Small teams are formed around projects in temporary figurations, with technical and non-technical experts added as required. Data analysts always work in tandem with "someone from business", to ensure that data work aligns with operational targets. As an escalation from other cases in the sample, the publication's data team views itself—not its data infrastructures such as data lakes or warehouses—as the "single point of truth" for all business-related data. Matching these aspirations, humans need to ensure consistency by glancing over the data every day, sharing ownership over its veracity and validity. The corporate data team works with external engineers, carries out the "light dressing up" of data, and approaches their work somewhat mechanistically as providing interfaces for both "machines and humans". In terms of metrics, C6 once again applies the mental model of the customer lifecycle, with units positioned along its various stages. Unique within the sample, we find a discussion around explicit data—as human agents are tasked with acquiring such customer data, that is provided voluntarily or explicitly. In addition to financial key performance indicators, the publication regards a composite metric of user engagement as the most important metric of all, as it correlates with likelihood to pay. Metrics are circulated around the organization through a common dashboarding interface, and overall, interviewees seem to be aligned in their use and understanding of metrics. As stated above, data workers at C6 focus their efforts on "serving" and "understanding" both internal and external customers, forming temporary team figurations. Practices involving quantification and correlation are less emphasized, yet the centralization of data work clearly normalizes interdepartmental cooperation on matters of data.

[154] See also 8.1, "Requirements engineering"

As demonstrated in 6.3.1, recently established data teams are predominantly led by domain experts external to the field, who then employ data engineers or scientists. While different parties "coalesce around the same discursive regime" (into an advocacy coalition), they pursue distinct agendas (Kitchin, 2022, p. 29) and operate in constant experimentation mode in their measurement of and optimization for metrics. A prevailing discourse on metrics observed throughout the sample suggests that optimization towards paying subscribers serves as the primary motivator for the adoption of data work. The notion of fully aligning with subscriptions seems difficult to challenge, gaining a lasting influence in practice, even if the discourse may ultimately prove to be unsustainable—subscription numbers will eventually plateau, knowledge will disseminate, and expensive data science (and machine learning) methods will be rendered obsolete (as anticipated by C6-4). As a result, the practices of data workers follow the principle of "making tangible" (*Sichtbarmachung des Vorhandenen*), providing evidence of growing numbers, where data or metrics constitute chains of command to varying degrees (with the regional publisher C4, perhaps surprisingly, an extreme example of metric ordinance). Data is continuously worked into evolving metrics aligned with changing goals. More specifically, data practices center around communication, whether in person or through dashboards and reports. In face-to-face interactions, the practice of "questioning the questions", and interpreting stakeholders (more common in sales and marketing, less so in editorial) frequently emerges in the sample. As other studies have found (Loukissas, 2019), experience of and a "feeling for" (Garnett, 2016) data is shaped by the data affordances available across the stream. In addition, the related practices of advocacy and enablement, key activities carried out by data teams, create looping effects controlled by the knowers of knowledge (Hacking, 2007, p. 306).

Notably, data imaginaries are at times represented by single metrics attributed with panoramic or prophetic qualities, like the controversial composite metric (ongoing experimentation with performance indexes at C3, scrapped altogether at C4) or the strategic North Star metric (C3). As opposed to the findings of previous studies, where experimentation happened as a consequence of "constant changes in audience structures and the web" (Belair-Gagnon & Holton, 2018), experimentation now becomes a function of the quest for correlations.

Overall, as data workers explore different approaches through trial and error, and specific types of metrics emerge from experimentation, it is reasonable to conclude that neither the metrics nor the strategies around them have stabilized, while the practices of individual data workers appear relatively homogenous across the sample.

6.3.3 Data transcending boundaries

As established in Chapter 2.2.3, boundary objects refer to artifacts, concepts, or tools that are flexible enough to be understood and used by different groups or individuals within an organization. In journalism studies, web analytics (referred to as editorial analytics in this study) are associated with the concept of interloper media or media interlopers (Belair-Gagnon & Holton, 2018; Holton & Belair-Gagnon, 2018), which refers to new media platforms or technologies that disrupt or challenge existing media structures and practices—in the case of web analytics, the companies who provide them. In the previous subchapters, I have established how editorial analytics are just one component within larger data assemblages, considered as one data source among many by data workers. Additionally, the technological stacks have become more sophisticated and include more than just the aforementioned analytics software provided by the interlopers. Based on these findings and in alignment with the corresponding guiding hypothesis outlined in Chapter 4.2, the aim of this section is to compile observations related to the following question:

Research Question RQ6

How would data affordances and objects work towards either transcending or reinforcing inter-departmental boundaries between the editorial and publishing domain or other demarcations?

In the case of C1, the establishment of the data department and the technical aspects of data affordances were predominantly driven by the publishing side of the business, as editorial "played no part there" (Interview C1-3). In addition, the editorial analyst sees editors as generally dismissive of data-informed thinking, possibly due to a perceived loss of control over the data narrative that editors used to command. On the other hand, the analyst considers data

mostly useless in the hands of editors, stating that "if editors didn't have data, the product wouldn't be much worse" (Interview C1-3). As in C1, we learn how editorial staff at C5 had no say in the establishment of the data department either: "The changes in journalism are of a structural nature and the publisher decides on these structures, not editors." (Interview C5-1).

With the two factions at C1 entrenched on either side of the data spectrum, certain metrics are intentionally withheld from editors (and work councils who might rebel against the metrics). Nonetheless, conversion data informs the placement of articles by editors, even though the causalities behind the metric remain unclear to them and are subject to exploration. Besides metrics displayed via dashboards that transcend boundaries and unify the departments under a common goal, reports could be qualified as boundary objects as editors and managers alike routinely request and read these reports C1. Similarly, a certain tension between editorial and the various data teams was noted in all interviews at C2. Specifically, the two groups optimize for conflicting goals, with subscriptions (strategic fencing off of content) pitted against readership (maximum public availability). On the other hand, the paywall is depicted as an object of collaboration between the publishing and the editorial teams (Interview C2-3).

Adding another layer of complexity to the discussion, the head of product at C3 views the firewall between editorial and publishing side in journalism as largely obsolete, asserting that data work has "no influence on journalism at all" (Interview C3-1). Grappling with a cognitive dissonance, the manager oscillates between framing data as insignificant to any and all journalists on the one hand, and insisting that journalists should wield data power over management on the other.

Could the erosion of the figurative firewalls between editorial and publishing in journalism be partly attributed to data work? At C5, the head of development sees interdepartmental divisions fading as the idea of temporary, cross-functional teams becomes more common (Interview C5-2). Another interviewee argues that collaboration across editorial and technical departments represents nothing less than a "paradigm shift in journalism" (Interview C6-2). While data work itself may not be the primary driver behind this refiguration of teams (with the focus instead being on the concept of holistic product design), the management framework Objectives and Key Results (OKR) is seen as instrumental in uniting personnel from editorial, product, and technology around a shared goal (Interview C5-2). On the surface, the narrative of fluid teams may seem attractive. However, these team constellations arguably shift autonomy away from individual departments to a metrics regime, with OKR or other "key" metrics guiding the discussion. In this way, definitional power moves to product or sales departments under the guise of the boundary metric (as explicitly stated by C6-1). Dashboards become the primary data affordance for accessing performance metrics, with their information potential depending on the user accessing them. An example of such "interpretive flexibility" (Star, 2010, p. 602) is evident when an analyst uses dashboards "on a more imperative, deeper level", enabling them to better "create new dashboards himself, [to] go deeper into analyses" (Interview C6-2) than an editor would. Overall, with editorial analytics fostering profit-oriented norms and becoming routinized in news production, the normative shift could now be considered as complete in two respects: a) the recreation of analytics infrastructure, dashboarding and reporting affordances, and b) the "making up of metrics" by the publisher, whereas these two aspects were previously controlled by outside interlopers.

As publishers start operating internal dashboards with proprietary data, this places the burden of providing "truthful data" onto the publisher as well—which had historically been a challenge for web analytics companies to solve (Belair-Gagnon & Holton, 2018, p. 8). In addition to these two lower-level boundary objects, the all-encompassing boundary object of news production could be considered as supplanted by subscriber production. Now that they are under the control of the publisher, these reconfigured boundary objects are not neutral either, as they have marginalized editorial power through renegotiation (see also Lee, 2007).

6.3.4 Under the data gaze

After discussing data practices across the sample and how specific data assemblages evolve around them, this section focuses on the concept of the data gaze at the individual level. As outlined in 2.2.4, Beer conceptualizes the data gaze in four parts; starting with a) the general assertion of analytics as a powerful data imaginary that informs the gaze, b) how temporality further shapes the gaze, c) the analytical spaces of codified clinics as a necessity for the gaze to flourish, and finally, d) the diagnostic eye of data analysts and engineers. At this juncture, I have to acknowledge the overlaps within my theoretical framework again—Kitchin draws on Beer in describing the data imaginary, while the notion of the codified clinic competes with the data assemblage, normative standards of a professional caste shape or (iso-)morph organizations according to their tastes and preferences. Further complicating matters, not all interviewees in the sample fit into the category of analysts or engineers (Beer limits his reflections on the diagnostic eye to this group). But, and this might be the first observation here, as I deliberately broadened my scope to include all individuals and functions involved in data work, we can see how Beer's dichotomy between data service providers carrying the data gaze and the organizations they supply falls short. Data affordances, workers, the creation of metrics, and in turn, new data practices, have migrated (or were drawn) into the organizations. Against this backdrop, my first research question emphasizes the subjectivity of data workers inside these organizations:

Research Question RQ8

How do data workers reflect on their own agency, potential conflicts of interest and the agenda setting power of their own work?

With the discussion around data assemblages in 6.3.2, encompassing the technical and contextual stacks, I have addressed what Beer refers to as the codified analytical spaces of the data gaze (Beer, 2018, p. 56). Analogous to medical clinics and Foucault's original conception of the emergent medical gaze, certain material infrastructures had to be established for the data gaze to become operational. In 6.3.2, I also discussed the rationality of speed, the pervasive trope of data flowing into dashboards almost instantly ("real-time"). However, I have also examined how this trope is often challenged by data workers in our sample.

In general, the expert interviewees tend to characterize their activities as passive or supportive. Whether consulting (C2-1), helping achieve results (C5-2), or imparting knowledge (C2-3), these individuals from middle or upper management do not directly acknowledge their agency. This type of understatement appears to be a common thread, with key personnel actively downplaying the role or sophistication of data in their organizations. For instance, the head of product at C3 remarks how the team often "just decides to make believe and move on" (Interview C3-1). Similarly, despite overseeing all data operations at C4, the head of data states how he "really can't do much, I am unable to generate or use data. But I can mediate and help stakeholders find a common language" (Interview C4-2). An editorial analyst (C1-3) tries to "keep away from interpretation" and thinks the trained data scientist is "able to really see things, not just dumb artifacts. I am unable to do that. Those people work in product, where more money is at stake." On the other hand, as the arbiter of dashboards, the analyst acknowledges his data power as a way of "nudging" editors towards certain decisions. In this sense, dashboards serve as a "backdoor to exert influence without telling anyone" (Interview C1-3) to the analyst.

Foucault argued that a new form of knowledge gives rise to a new type of measured language (Foucault, 2003, p. 114). In the pursuit of truth, the medical gaze seeks out an exhaustive description (Foucault, 2003, pp. 113–117) of its subjects—not of the totality of the human body, but of the details needed to prescribe treatments (Foucault, p. 100; Foucault, p. 196). In our sample, the quest for correlations (with desired subscriber behavior) represents this quest for the appropriate prescription. In their pursuit of correlations, managers at C2 seem to demarcate their specific data expertise with an extreme degree of measured language—the head of subscriptions (C2-3) uses sales metrics to establish the upper and lower bounds of a success corridor, while various heads of data take command of general data terminology such as "single point of truth" or "data warehousing". When articles "convert" and users "engage", behaviors are ascribed so off-handedly that we can assume they constitute standard vocabulary and form part of a "solidified jargon" (Beer, 2018, p. 112). However, interviewees also distance themselves from this language—calling user engagement a "useful buzzword" (Interview C6-1).

Another pattern emerges in the equation of data work to science and research, further reinforcing the notion of data as a source of truth and objectivity (a claim also made in the case of data-driven journalism; See also Splendore, 2016). The data researcher claims to operate with a purely "scientific and academic approach" to data, not following a daily routine in the sense of "looking at ten dashboards every day to check out the numbers. Instead, my work is more like implementing individual research projects" (Interview C5-1). Similarly, a head of innovation expresses their "admittedly soft spot for data due to [their] history in science" (Interview C5-3). As outlined in 6.3.1, data work at C5 takes place under the guise of "research", partly to appease editors, and it makes sense for staff to adopt this notion in their language as well.

A chief data scientist at regional publisher C4, reporting to a conspicuously humble head of data, aims to "always apply scientific methods to all questions of business" and go beyond "simply analyzing Excel spreadsheets and that's it". (Interview C4-3). In claiming scientific qualities, a demarcation is established between ostensibly innovative and superior data practices and previous approaches.

Analysts and other data workers are focused on unlocking the potential of data and establishing a data truth (Interview C6-3). Despite the numerous accounts of data work beginning with ad hoc requests from stakeholders in 6.3.2 (where data workers do not exclusively set the agenda), the analysts here maintain the right to "question the questions". Naturally, re-framing the problem space as the starting point of analyses influences the outcomes, and the negotiation power of stakeholders in these situations depends on their level of technical understanding. As one head of data acknowledges, "inside of marketing [...], there are colleagues, who are able to discuss with us almost on the level of SQL [155] and say precisely how they need a selection and even provide pseudocode[156]. Others are not so technically adept, coming from more creative areas. They naturally place a degree of trust into what we are doing" (Interview C6-3). In this way, expert knowledge on data engineering and analytics further solidifies an in-group around marketing and data, excluding editorial workers. Despite this, the researcher (C5-1) understands data work as fostering individual agency in the development of data and research practices, warning against the "fetishization of data". In sum, interviewees at four out of six organizations engage in these autoethnographic reflections on the limitations of data and data work.

[155] See also 8.1, "SQL".
[156] See also 8.1, "Pseudocode"

What can be said about the specific diagnostics carried out by data workers in our sample? There are two types of diagnoses identified by Beer, one conducted by analysts directly on the data, another to "inform, build and maintain the infrastructures" in which analysis occurs (Beer, 2018, p. 123). Typically, analysis falls within the domain of business analysts or data scientists, while maintenance is the responsibility of data engineers. In our sample, the roles are not always clearly defined and both diagnoses are carried out by fluid configurations—as exemplified by one head of data (C1-2) using the term "data persons" for her team comprised of both engineers and analysts. When speaking of their work, interviewees often evoke a mental model of processing systems, which aligns more closely with the qualities associated with data engineering. At C6, the group head of data (C6-3) sees his work as comparable to the journalistic production process, in his view a form of data processing. Similarly, another data manager (C1-2) understands journalism as a conveyor belt production process, implying the potential of measurability and optimization. In her role focused on "planning and maintaining the data warehouse", a head of data (C6-3) takes on a custodial role over the data infrastructure. Every morning, the codified clinic undergoes a quick visual inspection, a literal case of glancing at data: "We look at key metrics in the most important areas every morning and see if it all makes sense." (Interview C6-3) Again, more weight is given to operation and facilitation, keeping the codified clinic running, rather than the expert analysis of data. Building upon the findings as illustrated in 6.3.3, as diagnostics have moved from analytics companies to the publisher, the codified clinics of these companies have seemingly realized their "dreams of transcendence" (Beer, 2018, p. 133). However, the analytical capabilities of these clinics in the hands of their new practitioners remain ambiguous.

Beer describes how the "restlessness" of the data gaze (Beer, 2018, p. 128), its impulse to pursue total description, is central to the way that its knowledge is legitimized. In my sample, I see this restlessness reaching its limits in several counter-discourses of a) reducing the overwhelming quantities of data again (C3), b) the realization that data science may not always yield satisfactory outcomes (C4), or c) questioning the financial viability of data work in general (C6). In a similar manner to the restlessness of the gaze, Beer portrays the data imaginary as an expansive force with a "diamond tip, used for cutting, chipping, tearing, and opening the spaces into which it can expand" (Beer, 2018, p. 15). In the material, the data imaginary demands tangible progress in the hands of managers, while the underlying clinic personnel scramble to find correlations to the much-vaunted progress. Executives move in and out of their data gazing personas, at times informing the construction of metrics and data architecture (C2) or lobbying for their cause with performance indicators (C4). Indeed, while many subjective statements in the material align with the concept of the data gaze, there is contradiction and resistance as well. This suggests a varied and nuanced grasp of data inside the news organizations, with both extreme commitment and critical perspectives on the power and authority of data of data.

6.3.5 Towards uniformity?

As institutional isomorphism is a cross-cutting theme of this investigation, I will now discuss the findings established thus far in relation to potential inter-organizational patterns, similarities, or deviations that support either of the three isomorphisms proposed by DiMaggio & Powell.[157] Before delving into the specifics, some general uniformity across the organizations in our sample can be asserted. Among other aspects, in a) the establishment of dedicated data departments or units within the last <5 years, b) organizational structure, with data teams set up by upper management, reporting directly to product or sales, often operating within matrix structures, c) the establishment of complex and costly in-house data infrastructure by these teams, d) an influx of data professionals with non-journalistic backgrounds into the field, e) a metrics regime closely tied to the business model of subscriptions, and finally f) the concepts of conversion and the conversion funnel as pervasive across the sample. However, simply identifying these similarities does not explain the underlying causes. I want to discuss "why" and "how" these organizations align in terms of institutional coercion, mimeticism, and normativity. In short, evidence of all three isomorphisms is present, and I will discuss each one in order of their prevalence in the sample. With this subchapter, I address the following research questions:

Research Question RQ4

Why have news organizations adapted or enhanced their ways of working with data in recent years and can we find similarities in their origins and reasoning?

[157] See also 2.2.2

Research Question RQ5

What kinds of professional backgrounds (and normative influences) do data workers at the examined organizations have?

Research Question RQ7

How have new data-related roles emerged within the publishing side of news organizations and to what extent do these roles, with potentially differing norms and backgrounds, influence the interpretation and application of data in journalistic production?

Normative pressure can be found in the sense of data bureaucratization, where data workers at the news organizations establish a cognitive basis for their occupational autonomy in the form of metrics and measurement infrastructure.[158] As they professionalize in terms of knowledge and tooling, analytics competence moves inside of organizations. They a) recreate infrastructure, dashboarding and reporting affordances, and b) enable the "making-up of metrics" by the publisher, two functions previously performed by interloper companies (Belair-Gagnon & Holton, 2018). Additional normative pressure comes from data professionals with non-journalistic backgrounds and their motives like rationalization (journalism thought of as "assembly line work") or professional superiority through scientific expertise (siding with "science over Excel"). This introduces a new type of personality at the organizations—a methodical, number-crunching attitude that was not as pronounced in previous generations of data workers.

[158] Metrics as performative of strategic goals, see also 6.3.2.

In this sense, while the adoption of data and data work reflects the imaginations of measurability and effectiveness, their adoption may also be regarded as internal and external signaling to the labor force—performative data work as status competition.[159] A higher degree of professionalization drives institutional isomorphic change (in accordance with predictor B5): data professionals tasked with assembling entire teams recruit a specific cadre of data specialists who then architect systems, analyze metrics, and maintain infrastructure.

At C4, interviewees tell of a broader data ecosystem, a data collaboration among regional publishers. Anonymized data is shared among the participating parties and joint hypotheses formulated. Essentially functioning as a standards body, this collaboration applies a universal performance quantifier that exerts normative influence across the participating regional publishers. Another interview supports the observation, highlighting how publishers actively seek out knowledge transfer through informal exchange or associations like the one discussed above, because they are "all in the same boat" (Interview C5-2). Information and knowledge are shared (even more so than before), as the competitive landscape shifts due to digital platforms threatening core business models and direct access to user data (Nielsen & Ganter, 2018). This aligns with one of the organization-level predictors for isomorphism: "The greater the participation of organizational managers in professional associations, the more likely the organization will be, or will become, like other organizations in its field." (predictor A6)[160]

[159] One can expect these new professionals to hire likeminded people from their respective networks. For more on homophily in hiring decisions, see: Rivera, 2012; Rivera, Söderström et al., 2010

[160] Or, as DiMaggio & Powell put it, "models may be diffused unintentionally, indirectly through employee transfer or turnover, or explicitly by organizations such as consulting firms or industry trade associations" (DiMaggio & Powell, 1983, p. 151).

Evidence of mimeticism can be found in the adoption of concepts borrowed from Silicon Valley and ecommerce, such as OKRs and the North Star metric. In this sense, the field extends to organizations outside of the immediate news and journalism context, as journalism converges with social media platforms and ecommerce. As one interviewee explicitly stated, this mimeticism is grounded in the "promise of salvation in data, that it could tell you something". (Interview C1-3) Despite awareness of how data success stories are promoted by Silicon Valley, "people [...] like to reassure themselves with data and make this world more tangible, which they feel is always slipping through their fingers. And in the process you often forget to implement something concrete in your data work" (Interview C1-3).

While objectively true for all organizations in our sample, I could not identify ambiguity (predictor A4) in terms of organizational goals in the statements of interviewees. However, there persists an inherent duality between economic or technological incentives, as emphasized by data teams on the one hand and journalistic integrity on the other. Although there does not appear to be an alternative to the subscription economy, or at least alternatives are not discussed (predictor B3, "the fewer the number of visible alternative organizational models in a field, the faster the rate of isomorphism in that field."), data and data technology practices are explicitly qualified as experimental (C1, C2, C3 & C6). This observation aligns with another field-level predictor: "The greater the extent to which technologies are uncertain or goals are ambiguous within a field, the greater the rate of isomorphic change." (B4) These changes carry a ritualistic quality, as organizations with a visible commitment to data and data work also demonstrate their willingness to adapt and innovate in terms of technology. As journalistic organizations not only compete for journalists but a wider digital labor force, this encourages mimetic

isomorphism (in line with DiMaggio & Powell, 1983). Another driver of inter-organizational modeling, structural changes such as the establishment of data departments are publicly observable (and extensively communicated by C2, C4 & C5), whereas changes in policy and strategy are usually less apparent to the outside world.

Another institutional dependency that has been scrutinized by academics at length is the influence of social platforms, which serve as powerful intermediaries shaping the ways in which digital news businesses operate (Caplan & Boyd, 2018; Bell & Owen, 2017; Nielsen & Ganter, 2018). Surprisingly, the influence of Silicon Valley companies on data and data work, whether direct or indirect, was not widely acknowledged (with C1-3 as the exception). Instead, multiple interviewees (C3-1, C5-2) perceived ecommerce companies as more suitable blueprints than either dominant social platforms or singularly successful news brands like the New York Times.

Coercive isomorphism, characterized by organizations appearing increasingly homogeneous and organized around "rituals of conformity to wider institutions" (DiMaggio & Powell, 1983, p. 150) as a result of formal and informal political, structural, or societal pressures, is least pervasive in the sample. A prime example of political pressure, recent EU regulations on data, especially the General Data Protection Regulation (GDPR), have significantly impacted digital news businesses. These regulations impose an enhanced level of transparency and accountability for data handling, compelling news organizations to provide clear information on their data usage practices. Additionally, unambiguous consent must be obtained before any action on user data involving third parties, such as targeted advertising, may occur (Sanchez-Rola, Dell'Amico et al., 2019).

Regulations such as GDPR force digital news businesses to invest in robust data protection measures, often requiring significant capital investments in consulting and technology.[161] In this sense, regulation places additional pressure on data competence and professionalization. Despite the imminent prospect of more regulatory measures around data and privacy,[162] and open-ended questions about the critical challenges for data work, these government mandates were rarely discussed in the sample (C6).

Finally, the overall absence of claims related to increased efficiency or narratives about the results of data work seems noteworthy. This observation aligns with the notion that once an idea becomes widely accepted, its adoption may lead to more legitimacy, but not necessarily to increased efficiency: "The myths generated by particular organizational practices and diffused through relational networks have legitimacy based on the supposition that they are rationally effective." (Meyer & Rowan, 1977, p. 347) Arguably, it is precisely the uncertainty and experimental pressures prevalent in the studied organizations that suggest to their managers and personnel how data infrastructure and data competence are suitable (and relatively low risk) goals to pursue—not the idea of measurable efficiency gains resulting from increased data and data professionalism.[163]

---

[161] For instance, in order to react to individual requests, companies need to systematically gather and store transactional data. The "right to be forgotten", another important aspect of GDPR, can pose significant challenges to digital news companies as they need to remove user data from all systems under their control.

[162] In a preemptive move against regulations, browser vendors have pledged to disable third-party cookies altogether. Meanwhile, other regulatory devices like the ePrivacy Regulation (ePR), the Digital Services Act (DSA) or the Data Act will continue to strengthen consumer rights in terms of security, transparency and interoperability (EU Commission, 2017; EU Commission, 2022)

[163] Here we come full circle to the mode of operation of digital platforms we mentioned in our introduction, which tackle uncertainty and rapid change as "perpetual experiment engines" (Crawford, 2014) and approach society in beta-testing mode (Marres, 2017).

7. Conclusion

7.1 Findings and summary

Starting from the recognition that data and data work impact news businesses far beyond the newsroom, this thesis aims to understand how and to what extent data affordances and data thinking lead to organizational and structural changes and might even reshape professional boundaries within those organizations. Based on a comprehensive theoretical framework, these research questions were explored using a qualitative research design, supported by six case studies at news organizations in Germany.

Having established the motivation and significance of my research, I then developed a theoretical framework to guide the selection of methods and design of the study. Recognizing both organizations and individuals as units of inquiry, I drew upon critical data studies, organizational isomorphism, boundary objects, and the clinical data gaze as theoretical pillars. I followed up with a combination of clarifications on key concepts such as data, datafication, and knowledge work, aiming to establish a common understanding before delving into the empirical analysis. Additionally, I discussed existing and current research on topics related to data in journalism, identifying gaps and opportunities for further exploration. Finally, based on these preliminaries, I formulated a set of guiding assumptions and research questions to guide the empirical investigation. Next, I detailed the methodology employed in the study—the fundamentals of case study research and semi-structured expert interviews as the primary method of data collection. I elaborated on the selection of cases and research participants, the development and adaptation of the interview guidelines and the evaluation of my material in accordance with general quality criteria.

In the final chapter, I presented the findings of the study through individual case reports and cross-case comparisons within the theoretical framework and existing literature.

Within my findings, I conclude that fundamental changes in data practices, organizational structure, and management culture at the organizations took place over the last few years. In most cases, new roles have been introduced across the data spectrum, leading to the emergence of a new professional class of data workers (e.g. data scientists, data engineers or data analysts). These roles are (with the exception of C3) organized within newly established data departments that operate independently from editorial. Across all cases, the shift in data practices can be attributed to a re-orientation towards subscriptions and individual customers (as opposed to revenue from display advertising). Multiple consistent narratives illustrate how attention to and investment in data at these organizations can be traced back to a devaluation of traffic and reach.

While the structural changes are evident, as to the innovativeness of data practices in the digital era, interviewees appeared split on the nature and implications of said practices. Some perceive the shift to data work as largely an extension of traditional practices, supported by advancements in technology (Interview C1-3), while others argue that the scope of data collection and sophistication currently available represents a significant departure (Interview C2-1). Despite these varied opinions, a common theme emerged with data work in its current incarnation widely characterized as an evolution fueled by digital innovation, larger datasets, and improved tooling. However, the fundamental questions it seeks to answer remain unchanged.

Arguably, the specific utility of this newer iteration of data work remains limited for the time being: "That's why people work with data and like to reassure themselves with data and make this world more tangible, which they feel is always slipping through their fingers. And in the process you often forget to implement something concrete in your data work." (Interview C1-3, Pos. 288–291)

Considering the original "provocations" (Dalton & Thatcher, 2014) of critical data studies (CDS), I set out to 1) situate data regimes in time and space, 2) expose data in terms of whose interests they serve, and 3) illustrate the ways in which data are never raw. As a prerequisite for addressing these provocations, I examined my material through the lens of data assemblages, conceptualizing them as composed of both technical and contextual stacks. Beginning with the technical stack, data dashboards emerge as a ubiquitous technical concept across all cases. However, manual reports containing figures and statistics are more emphasized by data workers as the primary data artifacts circulated towards and reviewed by decision makers. The inclination towards static representations, often delivered ad hoc, on a per-request basis, might be perceived as an anachronism that contradicts technocratic imaginaries of self-service and automation. Regarding specific technical data practices, I find that predictive analytics using machine learning (ML) typically occurs in a non-automated manner—if ML is employed at all. Among the cases where such narratives were presented, C1 runs offline models, yet these models have not been deployed to production. Similarly, at C5, archival data is enriched with metadata through the application of ML.

On the practice of churn prediction, which involves calculating the likelihood of individual subscribers canceling their subscriptions using ML, such analysis may occur in relative isolation (C6). Alternatively, interviewees outright dismiss such prediction as pointless for digital assets, emphasizing how they apply ML to data from print operations only (C4). Overall, it is evident that ML is not frequently mentioned by interviewees, and even less so the potential of data and artificial intelligence.

Certain patterns emerge across cases in relation to software stacks, the array of infrastructure and software systems used to gather, operationalize, and deliver data. Notably, none of the case organizations operate data infrastructure on premise. Instead, they use various cloud computing and SaaS products, with Google Cloud Platform prevalent in terms of infrastructure and analysis across multiple cases (C1, C4, C6) and Amazon Web Services (C4) or Microsoft Azure (C2) utilized to a lesser extent. As an example of another concept around data, business intelligence software (as explicitly discussed at C2, C5) suggests a heightened level of operational efficiency through data in its name. Despite the marketing language surrounding specific software applications, a pattern of demarcation between two parallel software stacks emerges in our sample. Marketing and sales departments are shown to use more sophisticated (and more expensive) tooling compared to their editorial counterparts, who remain linked to the relatively older concept of editorial analytics instead (C1, C2, C5, C6). Overall the idea of editorial analytics seems primarily associated with editorial purposes, whereas data and visualizations for product, sales, and marketing teams are realized with a comparatively more complex technical stack. At the very least, business intelligence software seems to represent the idea of a new level of professionalism as compared to analytics (e.g. C2-1).

In terms of specific metrics, a more pronounced thinking in key performance indicators, or metrics and especially a hierarchy of metrics becomes apparent. Overall, even though interviewees express their focus on individual users and behavioural metrics, these individual metrics are often overshadowed by the potential of users as paying customers, determined by their quantifiable progress through so-called marketing funnels. A solidified thinking in users as quantities, more or less susceptible and optimizable via conversion, can be found across the sample. Measuring and inferring causalities from one or multiple conversion quotas or rates then becomes fundamental to the organizations. For upper management, the conversion metric is particularly meaningful, "because it's basically expressed in money." (Interview C3-1, Pos. 411–412). As the organization most interested in audiences, with membership data at the center of journalistic production, C3 interacts with their members as domain experts and collaborators, classifying members based on geolocation, interests, and other criteria (an interesting avenue for further studies). In other cases, metrics are used in different ways, even transcending their inherent retrospective quality as they serve as progress measurements towards an imagined future (C4). With this practice, the metric also carries a sense of planned obsolescence—valid only until the organization reaches its implied target quantity.

Regarding the contextual stacks, which are theorized as the other main constituent of data assemblages, business goals drive the application and interpretation of data, with an overarching push toward optimizing subscribers. Each case exhibits unique data practices and discourses. C1 displays a managerial focus on increasing subscribers, with an ecommerce-like approach to metrics. At C2, data is seen as a tool for financial control and operational steering, with the primary objective of increasing profit margins.

C3, by contrast, combines commercial conversion metrics with interviewees taking a critical distance to these metrics and discussing the dangers of overinterpretation of data multiple times. At C4, the organization's general digitalization is seen as accomplished, and working with data to optimize hyperlocal and special interest content is considered the next evolutionary step, with the goal of obtaining "full local authority over data" (Interview C4-1). C5 places greater emphasis on qualitative data work and research along the lines of the social sciences, while C6 again centers its data efforts on understanding and serving customers in the framing of a customer lifecycle.

Extending Ian Hacking's thinking, data here appears to not only contain prophetic qualities, as the specific quantifiers and metrics act performatively to shape the aspirations and outcomes of organizations or predict future events such as subscriber churn. Data also becomes inherently panoramic, as interviewees envision a golden record for every user, or establish the myth of a "single source of truth" as mentioned in multiple cases. Examples of powerful data imaginaries encapsulated in single metrics would be the controversial composite metric (ongoing experimentation with performance indexes at C3, completely abandoned entirely at C4) or the strategic North Star metric (C3). In contrast to the findings of previous studies on data work in journalism, where experimentation was driven by "constant changes in audience structures and the web" (Belair-Gagnon & Holton, 2018), experimentation now has evolved into a pursuit of correlations to a growth in digital subsribers. Overall, as data workers explore different approaches through trial and error and specific types of metrics emerge from experimentation, I have reason to believe that neither the metrics nor the strategies around them have stabilized, while the practices of individual data workers remain rather homogenous across the sample.

In terms of specific practices, the data team at C6 regards itself (not its data affordances, data lakes, or data warehouses) as the single point of truth to all business-related data. Notably, this terminology of a single point or single source of truth is at odds with the fundamental journalistic practice of using multiple sources to corroborate information before publication. Additionally, the practices of counting, quantifying, observing, reporting data and advocating for data work often converge into the single responsibility of the data teams. This creates a self-sustaining loop, although not in the narrow sense described by Hacking. The data team becomes a normative force of their own making. Other practices of data workers include making things visible, providing evidence of growing metrics, which are closely tied to varying degrees of managerial control (with C4 representing an extreme example). More specifically, data practices center around communication, whether in person or through dashboards and reports. In face-to face interactions, the practice of "questioning the questions" and interpreting stakeholders (particularly in sales and marketing, less so in editorial) emerges frequently in the sample. As other studies have found (Loukissas, 2019), experience of and a "feeling for" (Garnett, 2016) data are shaped by the data affordances available across the spectrum. Combined with the related practices of advocacy and "enablement", which are key activities carried out by data teams, another looping effect could be identified—as the "knowers" of knowledge (Hacking, 2007, p. 306) control both supply of and demand for data to some extent.

Overall, the original "provocations" of CDS surface across various organizations and how they practice data work. First, data regimes are situated in time and space through management culture, restructuring and powerful metrics. Each case displays unique data assemblages, with varying priorities for key performance indicators and emphasis on the automation of data work.

Second, the way data serves specific interests can be identified in how messages are constructed to align with corporate goals. For instance, metrics lean towards conversion rates and conventional financial targets, unifying the interests of sales and management in a single department (C2). Similarly in another case, management aims to achieve "full local authority over data" (Interview C4-1), further emphasizing corporate control and interests. Third, data here could never be considered raw as it goes through processes of ideation according to managerial interests (data originates from intentionally constructed tracking strategies according to intentionally created metrics), normalization in a technical sense (harmonization towards a "single source of truth", data passing through and displayed in specific software), and interpretation (data specialists reserving the right to question the questions addressed to them).

I also find some new nuances in the doxa between the publishing and production side of journalism. In multiple cases, interviewees explicitly state how editorial had no say in the establishment of the data departments (two cases) and how some metrics are kept away from editors (C1). Data dashboards, which are accessed and acted upon by editors, afford a degree of publishing power to the data workers tasked with constructing and maintaining them: "I present things in a dashboard in a way that I think makes sense. Yes, that's my back door for exerting influence without having to tell anyone." (Interview C1-3, Pos. 241–249) Apart from metrics displayed via dashboards that transcend boundaries and unify the departments under a common goal, reports could be qualified as boundary objects, as editors and managers alike routinely request and read these reports (C1, C5, C6).

Similarly, a certain tension between editorial and the various data teams was discussed in several interviews (three cases), with the paywall cast as an object of collaboration between publishing and editorial (C2). Others dismiss the conditions that historically led to the establishment of editorial firewalls as obsolete (C3) or see interdepartmental divisions fading as the idea of temporary, cross-functional teams proliferates (C5). To one interviewee, collaboration across editorial and technical departments represents nothing less than a "paradigm shift in journalism" (C6). However, these claims should be met with skepticism, as those making them are the beneficiaries of the change. On the surface, the narrative of fluid teams sounds appealing, but these team constellations arguably transfer autonomy from individual departments to a metrics regime with Objectives and Key Results (OKR) or other key metrics steering the debate. In that way, definitional power moves to product or sales departments under the guise of metrics. Dashboards become the data affordance to access performance metrics, revealing their potential on who accesses them. As an example of "interpretive flexibility" (Star, 2010, p. 602), and the negotiation power of stakeholders depending on their level of technical understanding, a data analyst might use dashboards on a more imperative level and go deeper into analyses than an editor would (Interview C6-2, Pos. 322–331).

With editorial analytics having fostered profit-oriented norms as routinized in news production (Petre, 2018; Tandoc & Thomas, 2016; Cherubini & Nielsen, 2016), the normative shift could now be considered truly complete in a) the recreation of analytics infrastructure, dashboarding, and reporting affordances, and b) the "making up" of metrics by the publisher, two functions which were formerly owned by external analytics and other software providers.

As publishers begin to operate internal dashboards equipped with proprietary data, the burden of providing truthful data, historically a concern of web analytics companies (Belair-Gagnon & Holton, 2018, p. 8) is shifting towards the publisher. Under the control of the publisher, these reconfigured boundary objects (e.g. dashboards, metrics, and reports) are not neutral either, as they marginalize editorial power through "renegotiation" (Lee, 2007). In addition, as analysts and other data workers are preoccupied with unlocking data potentials, their ability to re-frame the problem space as the starting point of analyses significantly impacts the outcomes.

Turning to the clinical data gaze as proposed by David Beer (2018), the dichotomy between service providers embodying the gaze and the organizations they supply appears inadequate. As stated above, data affordances, workers, and the practice of "making up" metrics, have migrated (or were drawn) into the organizations themselves. Contrary to the idea of powerful clinic personnel, our expert interviewees tend to characterize their activities as merely passive or supportive. Whether they are consulting, assisting, or sharing knowledge, these individuals from middle or upper management do not directly acknowledge their agency. In fact, understatement emerges as a common theme, with interviewees actively downplaying the role or sophistication of data in their organizations. Somewhat in line with Beer, some managers at these organizations (C2, C4) seem to demarcate their specific data expertise with an extreme degree of measured or "solidified jargon" (Beer 2018, p. 112). Another conceptual criterion introduced by Beer is a certain restlessness of the data gaze, characterized by a tendency towards "total description" (Beer, 2018, p. 128). In the sample, I see this restlessness reaching its limits in the counter-discourses of a) advocating for reducing the quantities of data again, b) recognizing that data science may not always generate adequate outcomes, or

c) fundamentally questioning the financial feasibility of data work. Overall, while several subjective statements in the material align with the concept of the data gaze, there is contradiction and resistance as well. More weight seems to be placed on the operation, facilitation and maintenance of the codified clinic (Beer, 2018), rather than expert analysis on top of data. As diagnostics have transitioned from analytics companies to the respective publishers, the codified clinics of these companies have realized their "dreams of transcendence" (Graham, 2004b)—yet the analytical capabilities in the hands of its new practitioners remain unclear. As the data workers here operate with a specific jargon and are specifically trained on data competencies (unlike previous generations of data workers such as SEO experts), they can be said to introduce a new logic or frame of reference to the realm of data work in journalism. If this qualifies as a fully developed clinical data gaze, I find debatable.

On the question of isomorphism, a certain degree of uniformity of outcomes can be observed across the sample. With the aforementioned influx of data professionals, primarily from non-journalistic backgrounds (C1, C2, C4, C6), data teams as envisioned by upper management were established. These teams typically report directly to sales-related functions or upper management (C2, C4, C5, C6) and often operate in new organizational configurations adjacent to matrix management structures (C2, C4, C5, C6). There are indicators of normative pressure in the shape of data bureaucratization, with data workers at the news organizations establishing a cognitive basis for their occupational autonomy through metrics and measurement infrastructure. Further normative pressure is applied by data professionals with non-journalistic backgrounds (and their requirements in terms of working material), driven by motives such as rationalization (perceiving journalism as assembly line work) or professional superiority stemming from claimed scientific expertise.

This introduces a new type of personality within the organizations, characterized by an organized number-crunching attitude that was not particularly pronounced in previous generations of data workers. In this sense, while data and data work embody imaginations of measurability and effectiveness, their adoption can be seen as a form of internal and external signaling to the labor force—data work as status competition. Working with and leveraging data also becomes a duty for the technologically ambitious manager type: "It is the job of managers like me, who come from this [technical] discipline, to signal they understand data and make transparent decisions. That wasn't always the case in the past because managers were basically still stuck in old ways of thinking." (Interview C2-2, Pos. 550–554) As another example of both normative and institutional isomorphic pressures, a regional publisher in the sample (C4) participates in a data association of news organizations. In this association, the former market competitors develop common success metrics, normalize and pool their data to gain insights. Participants must agree on interoperable data formats for this collaboration to succeed, leading to a certain level of exchange and equalization in terms of data competence in the process. Here, data becomes a unifying force across organizations, representing a shared resource among competing market participants.

There is some degree of mimeticism in the adoption of concepts borrowed from Silicon Valley and ecommerce, such as the management and goal setting methodology Objectives and Key Results (OKR), and the North Star metric. In this regard, the field extends to organizations outside of itself, into social media platforms and ecommerce companies. Aptly formulated by an analyst, this extension could be seen as testament to "an ongoing data megatrend" and data offering "a promise of salvation" (Interview C1-3, Pos. 286–287):

"Everyone knows these success stories from Silicon Valley, which were pushed by the PR departments of said companies, saying how they achieved great success with data. That's why people work with data and like to reassure themselves with data." (Interview C1-3, Pos. 287–291)

Coercive isomorphism, characterized by organizations becoming increasingly homogeneous and organized around "rituals of conformity to wider institutions" (DiMaggio & Powell, 1983, p. 150) as a result of formal and informal political, structural, or societal pressures, could be considered the least obvious in the sample. An example of institutional dependency is evident in the influence of social platforms, which serve as powerful intermediaries shaping the operations of digital news businesses (Caplan & Boyd, 2018; Bell & Owen, 2017; Nielsen & Ganter, 2018). Surprisingly, this institutional influence on data and data work was not acknowledged during interviews beyond the example above. Instead, multiple interviewees (C3, C5) perceived ecommerce companies as more suitable blueprints than either dominant platforms or singularly successful news brands like the New York Times. An example of political pressure is illustrated by recent EU data and privacy regulations, especially the General Data Protection Regulation (GDPR), which has had a significant impact on digital news businesses. These regulations impose an enhanced level of transparency and accountability for data handling. News organizations are compelled to provide clear information on their data usage practices and unambiguous consent must be obtained before taking any action on user data involving third parties, such as targeted advertising (Sanchez-Rola, Dell'Amico et al., 2019).

This study also demonstrates that experimentation with data and data infrastructure takes place regardless of company size. Furthermore, the particular data imaginary adopted by an organization bears no clear correlation (either positive or inverse) with editorial power. For instance, the sample includes a legacy newspaper (C2) with strong tendencies towards data centralization and concentration, while another legacy organization (C5) operates a decentralized research unit that prioritizes qualitative over quantitative data. Lastly, the overall absence of claims regarding increased efficiency or narratives detailing the impacts of data work is noteworthy. This observation aligns with another central tenet of organizational isomorphism, which suggests that once an idea becomes widely accepted, its adoption leads to legitimacy but not necessarily more efficiency: "The myths generated by particular organizational practices and diffused through relational networks have legitimacy based on the supposition that they are rationally effective." (Meyer & Rowan, 1977, p. 347) As demonstrated, data work serves as an appropriate descriptor for the diverse range of activities related to the interpretation, organization, and manipulation of digital data points. With the ongoing datafication across all units and functions within news organizations, increased prominence has been awarded to roles that involve working with, interpreting, and making decisions based on data. These roles do not just involve data scientists or analysts, as this study shows, but also a range of professionals from marketing to product managers on the publishing side. The analogy of "working" one thing into another, of fusing and merging material, seems appropriate. Here, engineers set up the instruments and provide the repositories of recorded measurements. Analysts then take these measurements and work them into invented metrics, while scientists try to predict and optimize these metrics in alignment with managerial goals.

Adressing the original research question, what could be a suitable definition of data work? Paraphrasing one of the interviewees (C2-3), data work encompasses a range of activities that facilitate the allocation of financial resources under conditions of uncertainty. In another sense (as exemplified by C4), data work encompasses the sum of all activities directed toward generating and disseminating performative metrics, which serve the interests of the managerial staff.[164]

I hope this thesis serves as a meaningful contribution to the field of journalism studies by revealing how power dynamics in the news media industry are reshaped through the integration of data technology, data practices, and a new professional class of data workers. Despite efforts to downplay or deny these changes, the influence of this emerging professional class seems palpable and manifests in various ways, ranging from their capacity to shape metrics, a discursive power that defines successful journalism to their direct line of communication to and from upper management. All the while, the measurable impact of their work remains unclear. Future research would need to determine whether the snapshot captured in this study merely represents the early stages of data work and whether such endeavors will ultimately lead to increased efficiency.[165]

[164] Here we circle back to one of the introductory quotes: "What gets measured gets managed—even when it's pointless to measure and manage it, and even if it harms the purpose of the organization to do so." (Barnett, 2015)

[165] "I think that's the next step you have to take. To really work with the data. Not to fall into what they call a cargo cult. To think that just because you look at data, you're doing well." (Interview C1–3, Pos. 436–442)

7.2 Limitations and further research

In Chapter 5.2, I alluded to potential shortcomings in my research design. In addition to these methodical limitations, I aim to provide a comprehensive overview of the empirical limitations, inherent omissions in my material or interesting junctures I encountered over the course of the study and point out desirable paths for further research.

First, I must acknowledge the methodical limitations inherent in sampling expert positions across German news organizations through exploratory case studies. While this design provided valuable insights into these specific entities (including both legacy and digital native organizations of various sizes), it restricts the ability to generalize the findings to all news organizations, particularly those located outside of Germany. Furthermore, the utilization of semi-structured expert interviews, while effective at gathering nuanced information, introduces potential limitations of subjective interpretation and inconsistency in responses. Given the conversational nature of these interviews, the questions varied slightly between interviews, which may have influenced participant responses in different ways. This could impact the consistency of gathered data, potentially affecting its reliability and comprehension. Despite these limitations, the overall research design was considered appropriate for generating an in-depth understanding of the complexities of data work in news organizations while simultaneously exploring the phenomenon broadly. During the analysis phase, one has to acknowledge how potential researcher bias and interpretation (such as focusing on specific topics in the conversational tree, possibly overlooking cues in the material, and inventing codes and latent themes) might have influenced the results to some degree.

As for the scope, the study could be considered limited both geographically (restricted to Germany) and historically, as it represents a snapshot of data thinking before the widespread popularity, distribution, and discussion of generative artificial intelligence (AI/ML) software (towards late 2022).[166]

In turn, these limitations offer perspectives for further study. For example, while this study provided insights into data work within for-profit news organizations, it raises questions about how public broadcasting organizations approach data work, especially in the absence of economic pressures related to conversion and subscriptions. Future research might cross-reference studies from other domestic markets, especially (but not limited to) the Nordics, to gain comparative insights.[167] Having encountered a type of data trust or informal data standards body among news organizations suggests that such normative forces in data interoperability could be subjects for further investigation and how such research could be approached from both the avenues of organizational sociology and computer science, framing it as a standardization problem.[168] Additionally, while this study primarily looked for answers outside the newsroom, recent research around the adjacent concept of editorial technologists, considered as "engineers of sociotechnical change" in newsrooms, could be discussed in light of the findings here (e.g. Lischka, Schaetz & Oltersdorf, 2022).

---

[166] A popularity largely prompted by the release of the generative text model ChatGPT 3.5 by the US-based company OpenAI in November 30, 2022 (OpenAI, 2022).

[167] Numerous works have observed the remarkable agility of Nordic news organizations in adapting to digitalization and shifts in news consumption behaviors, for example: Lindberg, 2023; Franklin, & Eldridge, 2016; Anderson, Downie & Schudson, 2016. The ongoing Finnish project "The Future of Dispersed Journalism" (Ahva & Ovaska, 2023) explores data discourses beyond the confines of the newsroom.

[168] Specifics about the data trust withheld here for reasons of anonymity but gladly given on request.

Even though AI/ML and its promises were widely discussed at the time of sampling, there is a notable absence of reflections on AI/ML and other types of "cutting-edge" technology in the sample. This might be attributed to the fact that the interviewees were practitioners in the field—less inclined to to talk about future trends as opposed to pundits or journalists. If this study were to be repeated, the result would probably look different following the widespread discourse on generative AI/ML models. An early indication of this shift can be seen in recently published works examining the impact of text generation on news production (Pavlik, 2023), as well as the meta-discourses in the field (Moran & Sheikh, 2022).[169]

With the advent of generative AI/ML models, trained on vast portions of the public internet, news articles are effectively turned into inputs for automating journalism itself.[170] This raises concerns about embedding biases or falsehoods in journalistic production. What legal frameworks exist to either prevent the unauthorized training of generative models on proprietary data from news organizations or to enable these organizations to monetize this opportunity? Additionally, the potential of artificial intelligence to produce "deepfakes"—highly convincing counterfeit images, audio recordings, and videos—raises significant concerns regarding the epistemic integrity of data utilized in journalistic reporting (Chesney & Citron, 2019).

[169] Moran and Sheikh (2022) identify two primary discourses in a) the economic optimism vs. professional skepticism divide, where newsroom leaders and funders see AI as a cost-effective tool, while journalists worry about its impact and b) the lack of technical understanding among journalists, which results in narrow discussions around content production characterized by a sense of technological determinism.

[170] Only briefly discussed in C5, news data (structured or unstructured) held by the organization was not used commercially as training data for machine learning. To the author's knowledge, none of the organizations were selling training data at the time of sampling.

Another temporal aspect highlightes by interviewees (C1-3, C3-1), data work in journalism is evolving at such a rapid pace that this study risks becoming obsolete upon publishing.[171] Although research on editorial analytics may now be considered historical, this conondrum of studying contemporary phenomena was addressed by media researchers elsewhere and does not diminish the relevance of a single study as a reference point.[172] Instead, future research could explore the significance and perceptions associated with the concept of data science in news organizations again over time. One interviewee suggested that expensive data scientists would soon be replaced by service providers as organizations transition from the in-house experimentation phase of their data work.[173] Assumptions like these could be monitored and validated through longitudinal or panel studies.

[171] To mitigate this problem, I tried to limit discussion of specific technologies or market players in my analysis (assuming that specific tools are less relevant than overall organizational goals or imaginations) while keeping some minute technological details in the individual case reports.

[172] As an example, Bruns & Highfield (2016) discuss the speed at which social media platforms evolve, underlining the difficulty of capturing a "snapshot" of these platforms. Venturini & Rogers (2019) reflect on the speed of changes in API environments and the challenge researchers face when their findings become obsolete by the time they are published.

[173] "Data science appears to be the absolute non-plus-ultra. But calculating models and such are things that can soon be completely outsourced technically. [...] There are providers out there who can basically build scoring models within a few hours. Data scientists will be surprised to see how quickly they can be rationalized away, to be honest. [...] We do everything manually now and I think it's great because we get to know our company really well. But I don't think we'll need it any more in the long term." (Interview C6-2, Pos. 732–776)

8. Appendix

8.1 Glossary

Accelerated Mobile Pages (AMP)

Accelerated Mobile Pages are a suite of tools to effectively make mobile websites load faster by preloading content directly hosted by Google. It was introduced to facilitate access to content on mobile devices, particularly for users with slower internet connections. AMP has also contributed to Google's strategic goal of increasing engagement with their products by serving news article detail pages from Google servers. This led to a reduction in autonomy for publishers, as their content is effectively mirrored by Google.

Adobe Analytics

A software service which aims to assist in the gathering, organization, and understanding of so-called "analytics" data garnered from a broad spectrum of sources. Advertises the function of analyzing user activity on websites and across digital platforms, providing insights into customer behavior and preferences.

Aggregation

In the context of this study, *aggregation* refers to the process of summing up data points across arbitrary dimensions. For example, "subscribers" alone may not provide a complete metric; it needs to be measured across another dimension such as time (e.g. days or weeks). The smallest available measurements across time are then "rolled-up" or "aggregated". Due to the computational expense involved in calculating these time-aggregated metrics, they are often precomputed in advance.

Alerting

In computer systems, *alerting* involves systems sending out notifications in response to certain predefined events or situations. These notifications, known as alerts, are designed to draw attention to potential issues such as breaches in system security, failures in hardware, or other technical difficulties. They can be sent through various communication channels, including email or text message.

Amazon Web Services (AWS)

One of the major cloud computing platforms, operated by Amazon. It primarily offers time-shared and virtualized, remote computing services designed to support online and web-based applications. AWS provides tools for computing, storage, database, analytics, networking, mobile, developer, management, IoT, AI, security, and application services.

Application Programming Interface (API)

A set of rules and protocols for building and interacting with software applications. It defines the ways in which software components should interact, facilitating communication between different software programs. It can be viewed as a bridge between different software applications, allowing them to work together.

Microsoft Azure

Azure is the equivalent of Amazon Web Services, in this case operated by Microsoft. See also *Amazon Web Services*.

BigQuery

A software service provided by Google that facilitates the handling and analysis of large-scale data. It enables users to analyze extensive data sets in "real-time" using a SQL-like language (see SQL).

Business Intelligence (BI)

Commonly referred to by its acronym, BI encompasses various processes for analyzing and utilizing data to inform business decisions. BI is typically supported by software tools that extract and transform data, as well as provide interactive dashboards and visualizations for analysis. See also *Extract, Transform, Load*.

Customer Acquisition Cost (CAC)

Refers to the amount of money a company spends on marketing and sales activities to acquire a single new customer. This includes expenses such as advertising campaigns, promotional materials, sales team salaries, and any other costs associated with attracting and converting potential customers.

Churn

A business term used to describe the rate at which customers stop doing business with a company. It is often used in the context of subscription-based services, such as cable TV or mobile phone plans. In the field of *journalism*, churn would refer to the percentage of customers cancelling their news subscriptions during a given period, such as a month or a quarter.

Click-Through-Rate (CTR)

A quantifiable metric often used to express the performance of a digital offering, most commonly display or banner ads. For example, a banner that

was displayed a thousand times but clicked on only 10 times would have a CTR of 1%.

Consent Management

A process most commonly found on websites and integrated within apps which aims to ensure individuals are informed about and let them control how their personal information is collected, stored, and used by organizations. Under EU jurisdiction, website owners must implement and manage user interfaces that enable transparent consent from individual users regarding how their data is used and/or shared with other parties.

Conversion Funnel

Originating from the term *purchase funnel*, which stems from a marketing model first developed by advertising strategist Elmo Lewis in 1898 (Strong, 1925; p. 349f), who charted the hypothetical route of a consumer from the point of initial awareness of a brand or product to the moment of making a purchase. Later adapted by both marketing consultants and scholars, the funnel serves as a mental model for optimizing various steps along this journey. Similarly, the *conversion funnel* emerged from ecommerce operations, where it describes the route a buyer takes when navigating a digital shopping application before finalizing a purchase.

Conversion Rate

A statistical metric measuring the percentage of success for a digital offering or incentive to lead to a specific desired action. For example, if the option to sign-up for a digital newsletter was displayed a thousand times to readers of a news website but only 10 of these readers entered their email address, then the

newsletter sign-up in its specific form has a conversion rate of 1%. See also *Conversion Funnel.*

Cost-per-interest (CPI)

A financial metric that determines the amount of money spent on an advertisement or marketing campaign relative to the generated interest, usually measured by engagement activities such as clicks or views.

Cost-per-mille (CPM)

A pricing model in digital advertising where advertisers are charged a fixed amount for every one thousand views or "impressions" (loading and display process) of their advertisement.

Cost-per-order (CPO)

A statistical metric representing the average amount of money spent per order or sale generated by a specific advertising action (or "campaign"). It is calculated by dividing the total advertising spend by the number of orders generated by the campaign. Tracking the CPO allows businesses to assess the effectiveness of different channels, messages, and audiences. Not a metric per se, the *Max.-CPO* mentioned by some interviewees defines an *upper limit* on the maximum acceptable or economical CPO. This maximum is predetermined and serves as a benchmark for campaign performance.

Cost-per-response (CPR)

A statistical metric indicating the total cost of a marketing campaign divided by the number of responses or actions it generated, such as purchases or clicks, providing a measure of the campaign's cost-effectiveness.

Crawler

Also known as a *web crawler* or *spider*, a crawler is a piece of software that systematically browses the World Wide Web for the purpose of web indexing (web spidering). It visits websites and reads their pages and other information to create entries for various intents and purposes, for example, to build and update a *search engine* (Page & Brin, 1998).

Customer Lifecycle

Refers to the progression of steps a customer undertakes when considering, purchasing, using, and maintaining her loyalty to a product or service. The term encapsulates each stage a customer may encounter in their interaction with a company, including initial awareness or interest, making a purchase, using the product or service, and potentially establishing repeat business. See also *Conversion Funnel*.

Customer Relationship Management (CRM)

CRM is the "strategic process of selecting customers that a firm can most profitably serve and shaping interactions between a company and these customers. The ultimate goal is to optimize the current and future value of customers for the company." (Kumar & Reinartz, 2018)

Dashboard

A graphical interface designed to present and organize a collection of data in a (theoretically) coherent and easily understandable format, often used for the purposes of summarizing, analyzing, or managing relevant information in "real-time", providing users with insights into current events or changes as they occur.

Database

A self-describing collection of integrated records. A record is a representation of some physical or conceptual object. Self-described here means that the database should contain a description of its own structure, usually thought of as *metadata* to the data. The database is *integrated* in that it includes the relationships among data items, as well as the data items themselves (Kroenke & Auer, 2007).

Data Cube

A multi-dimensional (hence cube-like) representation of data, often used in business intelligence and data warehousing, facilitating the simultaneous viewing, organization, and analysis of multiple data points along different variables or dimensions. See also *Business Intelligence, Data Warehouse.*

Data Lake

A technological structure or system predominantly utilized in big data analytics, where vast amounts of raw data, heterogeneous in nature and from multiple sources, are stored in their native format until needed. The term *lake* signifies how data is not yet organized, taking on "liquid" characteristics. The more organized variation of a central (big) data storage system for data from multiple sources would be the *data warehouse.*

Data Warehouse

A structured repository of historical data, yet independent of a specific application, but already optimized for reporting and analysis. As opposed to the *data lake,* the "warehouse" was designed with a predetermined schema, meaning data is pre-cleaned, pre-aggregated, and structured.

Decision Tree

A tool utilized in machine learning (ML) and data mining. Mimics human decision-making abilities by selecting, categorizing, and analyzing data to predict probable outcomes. It starts at a single point, known as the root, which splits into branches that eventually end in leaves, representing outcomes.

DevOps

An approach in software development that integrates development ("Dev") and information technology operations ("Ops"). Its goal is to shorten the system development life cycle and continuously deliver high-quality software. Unlike the traditional waterfall methodology, where software is developed, shipped, and then evaluated as a whole, DevOps promises faster identification and resolution of problems by continuously delivering software updates.

(User) Engagement

A quality of human experience characterized by the depth of an actor's investment when interacting with a digital system (O'Brien, 2016). While mostly intended as directly correlating with conversion (to a digital subscription) in the context of this study, the ability to engage and sustain engagement in digital environments can also result in positive outcomes in, for example, citizen inquiry and participation, electronic learning and so on. In the sense of human-computer interaction, the abstract construct of user engagement (UE) manifests differently within different computer-mediated contexts, which has made it "challenging to define, design for and evaluate" (O'Brien, Cairns & Hall, 2018, p. 28).

Extract, Transform, Load (ETL)

Refers to the general data integration process comprising three main steps: a) the *extraction* of data from multiple, often disparate, databases or other sources, b) the *transformation* of the data into a uniform format, followed by c) its *loading* into a destination system such as a data warehouse or a data lake. See also *Data Warehouse & Data Lake*.

Event Analytics

In this case, an "event" refers to a specific action or interaction which is recorded or tracked on a website, app, or other digital affordance. It could be triggered by a user interaction or behavior, such as clicking a button, making a purchase, watching a video, or filling out a form.

Event Pipeline

A system where streams of digital "events" (incremental point-in-time data) are processed through a series of stages, with each stage handling specific tasks such as filtering, transformation, aggregation, or enrichment. One example is an event stream processing system used in financial trading, where incoming market data events are filtered, analyzed for patterns or triggers, aggregated for trend analysis, and finally output for decision-making or alerts, often in dashboards.

GDPR

Acronym for "General Data Protection Regulation", a legal framework instituted by the European Union in 2018. It sets standards for data protection and privacy for individuals within the EU and European Economic Area and regulates how businesses and public organizations handle personal data (EU Regulation, 2016).

Google Cloud Platform

A major cloud computing platform operated by Alphabet/Google, see also the equivalent (to an extent) *Amazon Web Services* or *Microsoft Azure.*

Graph Theory

A branch of mathematics concerned with the study of graphs; mathematical structures used to model pairwise relationships between objects. It involves various aspects such as analyzing, interpreting, and understanding the properties and behavior of graphs (Trudeau, 2013).

Header Bidding

A programmatic advertising process where publishers offer digital advertising space to multiple digital advertising exchanges simultaneously before it is filled with an advert. This allows all potential buyers to compete with each other, theoretically leading to better yield for the publishers.

HTTP Status Codes

Three-digit numbers that are returned by servers to indicate the status of a web request made through the *Hypertext Transfer Protocol* (HTTP). These codes are divided into five classes, where the first digit identifies the class. The codes offer information about the success, failure, or other status of the request. The *Internet Assigned Numbers Authority* (IANA) maintains the official registry of these codes.

Hypertuning

Refers to the process of optimizing machine learning (ML) model parameters to improve its performance or prediction accuracy. Notable methods employed

during this process include *Grid Search* and *Random Search*, effectively aiding in the selection of the most optimal hyperparameters for any given model.

Kanban

"Kanban" ("visual signal" or "card" in Japanese) is an industrial manufacturing methodology which originated in Japan to effectively schedule inventory levels and production processes. However, Kanban was later adopted by software development as an alternative to the traditional "waterfall" methodology (doing things in strict sequence, according to a plan, after an extensive planning phase, delivering only completely finished products). In this context, Kanban signifies an approach to software development that emphasizes flexibility and so-called "continuous delivery". The rationale of this approach is to improve efficiency by only working on what is necessary at any given moment ("just in time") and frequently delivering small increments of functionality.

Key Performance Indicator (KPI)

A metric used to measure and evaluate the success or efficiency of a project, organization or individual against their set goals or objectives.

Leads

A marketing term for individuals or organizations who have shown interest in a product or service, suggesting potential for a future sale. Leads can be generated through various channels like direct inquiries, referrals, or marketing campaigns. As they "convert" from one stage to another, leads can evolve into more "qualified" leads, meaning the contact was evaluated and its potential further substantiated or contextualized by one party or another. (Steenburgh & Avery, 2010).

Logfiles

Text files generated by computer systems that record operations and transactions performed within the system. Essentially a form of documentation, the act of "logging" chronicles error messages, user access, system start and shutdown data, incoming or outgoing requests (transactions over the internet), and other arbitrary activities. These records are critical for system maintenance, troubleshooting, auditing individual software, and/or security reviews. No single uniform standard for logfiles exists as their form and content vary by system or application type.

Machine Learning

A subset of artificial intelligence, involving software designed to learn from data or experience to improve its performance, prediction, or decision-making capabilities "without being explicitly programmed" (Samuel, 1959). Machine learning is based on the concept that machines can learn patterns, adapt their understanding, and make intuitive judgements similar to humans. Another often cited definition: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." (Mitchell, 1997)

Metrics

Arguably a popular term for "applied econometrics" (Angrist & Pischke, 2015), "metrics" more broadly refer to any form of measurement used to gauge some quantifiable component of a performance. But whereas the field of econometrics applies statistical methods to economic data for empirical analysis or testing economic theories, metrics can be *any* quantifier for *any* purpose of comparison and evaluation.

Natural Language Processing

A field of study in artificial intelligence and computational linguistics that focuses on the interaction between computers and humans through natural language. Its primary goal is to design algorithms and models that allow computers to comprehend (also known as natural language understanding or NLU), interpret, and generate (also known as natural language generation or NLG) human language in a valuable way. Pioneering a mathematical theory of language syntax, later extended to a whole "generative grammar", Noam Chomsky illustrates language as a system of syntactic structures which would "provide an account of a hypothetical language-learning device and could thus be regarded as a theoretical model for the intellectual abilities that the child brings to language learning." (Chomsky, 1966).

Objectives and Key Results (OKR)

A goal-setting framework used by organizations, consisting of relatively specific overarching objectives and underlying quantitative thresholds or subgoals to track progress towards achieving the objectives. OKR originated at Intel in the 1970s and was popularized by Google in the 2000s.

Paywall

A digital barrier created by a news media organization that restricts access to their online content to specific users or subscribers. Such a barrier can be implemented in various ways, either limiting the number of articles available freely (without a log-in or subscription) or blocking all access to content until a subscription is purchased. A *metered paywall* would offer a limited number of articles for free (in a moving monthly or weekly window) whereas a *hard* paywall would require payment to access any content at all, and a *soft paywall* restricts access to select articles only.

Performance Marketing

A form of digital advertising where advertisers only pay if specific actions or targets are completed or fulfilled (such as clicks, conversions, or sales), aligning the cost of advertising with its measurable level of impact. See also *Conversion*.

Predictive Analytics, Predictive Learning

A branch of advanced analytics, which utilizes data, statistical algorithms, and machine learning techniques to determine potential future outcomes based on historical data. *Predictive Learning* is an adjacent technique in machine learning employed across various disciplines, ranging from weather forecasting to stock market trends.

Programmatic Advertising

Automated buying and selling of digital advertising inventory, leveraging technology in the form of algorithms and agents which negotiate based on target audiences and demographics. Aspects of programmatic advertising include specific processes such as real-time bidding (RTB) and targeting or re-targeting (Busch, 2014).

Pseudocode

A way of representing a computer program using plain language not specific to any particular programming language. Pseudocode helps programmers plan and outline the logic and structure of a program before actually writing it in a specific programming language.

Requirements Engineering

A systematic process in systems and software engineering that involves defining, documenting, verifying, and managing stakeholder requirements. It

encompasses steps such as identification, elicitation, analysis, specification, validation, and monitoring of these requirements (Demarco, 1986). It establishes a bridge between system stakeholders and system developers, balancing their needs and potential trade-offs.

Retention

An inverse metric to the metric of churn. Expressed in percentage (the rate of retention). See also *Churn.*

Search Engine Optimization (SEO)

A spectrum of practices and techniques used to enhance the visibility and ranking of a web page in search engine results (e.g. Google or Bing). The primary goal of SEO is to gain a higher volume of web traffic by enhancing the page's content relevance to specific search terms (also referred to as *editorial* SEO). SEO also involves ensuring that the page adheres to other guidelines, technical evaluations and criteria set by search engines such as accessibility and performance (also known as *technical* SEO).

SHAP Values

The resolved acronym of "SHapley Additive exPlanations" (in honor of mathematician Lloyd Shapley), SHAP values are a unified measure of feature (an individual measurable property or characteristic of the observed data) importance in machine learning algorithms. They assign each feature an importance value for a particular prediction, based on the contribution of each feature to the prediction's uncertainty. Derived from game theory, SHAP values are intended to be fair, consistent, and locally accurate attributions that sum up a prediction in its totality.

SQL

Structured Query Language, a programming language developed in the 1970s by IBM researchers Raymond Boyce and Donald Chamberlin to manage and manipulate relational databases. SQL provides a standardized way to interact with databases and has since become the de-facto standard for interrogating large datasets.

Support Vector Machine

A type of supervised Machine Learning (ML) model primarily used in classification and regression problems. It operates on the principle of maximizing the margin around the separating hyperplane in a high- or infinite-dimensional space, striving to find the optimal separating hyperplane that maximizes the distance between data points of different classes. The data points that touch or define the margins are called support vectors. According to Cortes and Vapnik, the "support-vector network is a new learning machine for two-group classification problems. The machine conceptually implements the following idea: input vectors are non-linearly mapped to a very high-dimension feature space. In this feature space a linear decision surface is constructed" (Cortes & Vapnik, 1995, p. 273). See also *Machine Learning.*

Survival Analysis

A statistical approach primarily used in the medical field, the social sciences, and engineering disciplines (as in the context of this study), involving the analysis of *time-to-event* data for events such as "death" in biological organisms or "failure" in mechanical systems (Lawless, 2011). Survival analysis looks for situations where the final event (death or failure) has not occurred for several subjects during the period of observation and measure the risk of the event occurring over time.

Tags / Tagging

In the context of this study, tags, or the practice of tagging, refers to the manual or automated generation of metadata for articles or other content. A tag here is a label or a piece of information that further classifies, groups or denotes an item with the intention of increasing searchability, navigation and discoverability of the item. Widely used and operationally relevant tags in news organizations are, for example, general sections or topic keywords.

8.2 Interview Guideline

Final iteration, translated from German

Introduction & gratitude

- "I would like to start by briefly explaining why I wanted to conduct this 60 minute interview with you..."
- Focus of investigation: What does working with data look like in journalistic organizations beyond the newsroom? If any, which roles and ways of working with data are emerging?
- Aim to understand the changes in data work against the backdrop of continued digitalization.

Module 1: Personal background & role perception

- Please tell me about your position, title and education.
- What do your daily tasks typically look like?
- How would you describe your area of responsibility within the company?
- Where do you place yourself within the company (the organizational chart)?

Module 2: Individual forms of data work

- How would you rate the role of data and metrics in your work?
- Which specific data and metrics are important to your work? Please give examples if possible.
- Who provides the data relevant to your work?
- What tools or applications do you use in connection with data or metrics?
- Dashboards & visualizations

- Do you use dashboards or similar interfaces?
- If so, can you describe these dashboards in more detail for me? Might there even be public material on these?

- What role do external collaborators or service providers play in terms of your data work. Can you provide examples?
  - Cloud service providers like *Amazon Web Services*
  - Consulting firms
  - Startups

Module 3: Forms of data work in the organization

- On specific databases, -systems and their purposes
  - Where are data, databases or metrics particularly important inside your company? If possible, provide examples.
  - If at all, to what extent has your company's way of working with data changed in recent years? If possible, provide examples.
- Interdepartmental cooperation on data and data systems
  - How important are editorial requirements in the planning of data systems?
  - Are there datasets or -systems which are less relevant for editorial work? Please elaborate.
  - How do you assess the influence of different business units on data work in your organization?
- Specific characterizations of the relevant teams and roles
  - In terms of data, have new roles have emerged recently? What are their contributions?
  - Have any roles disappeared in the process?

    - Who is involved in the procurement of data? Who is involved in the analysis?

Module 4: Data work in general and look into the future

- Were there structurally similar databases and/or -systems in the past or does working with data in journalistic organizations now represent something inherently new?
- How would you rate the influence of other companies (e.g. *Google*, *Netflix*, *The New York Times*) on working with data and metrics in journalistic organizations overall?
- If you were to imagine, how might journalistic organizations work with data and metrics in five to 10 years?
- Generally speaking, what are particular challenges in dealing with data and metrics?

Module 5: Conclusion and reference procedure

- Would you like to add anything else on the subject? Have we glanced over important aspects here?
- Would you be able to recommend other people in your area to talk to? Can you establish contact?
- Thank you for the conversation
- Clarification on anonymization & quotes

Postscript / Notes on the interview

- Impression of the interviewee (level of engagement, mood, cues etc.)
- Place, date & length of the interview / context of the interview
- Notable occurrences during conversation (interruptions, problems)
- Notes on epilogue after the recording, if applicable

8.3 References

Abbott, A. (2014). *The system of professions:* An essay on the division of expert labor.

Ahva, L., & Ovaska, L. (2023). Audience metrics as disruptive innovation: Analysing emotional work of Finnish journalism professionals. *Nordicom Review*, 44(2), 152–171.

Akkerman, S. F., & Bakker, A. (2011). Boundary crossing and boundary objects. *Review of educational research*, 81(2), 132–169.

Alaimo, C., & Kallinikos, J. (2022). Organizations Decentered: Data Objects, Technology and Knowledge. *Organization Science,* 33(1), 19–37.

Alfter, B. (2016). Cross-border collaborative journalism: Why journalists and scholars should talk about an emerging method. *Journal of Applied Journalism & Media Studies*, 5(2), 297–311.

Alvesson, M., & Spicer, A. (2019). Neo-institutional theory and organization studies: a mid-life crisis? *Organization Studies*, 40(2), 199–218.

Ananny, M., & Crawford, K. (2015). A Liminal Press: Situating news app designers within a field of networked news production. *Digital Journalism*, 3(2), 192–208.

Anderson, C.W., Downie, L., & Schudson, M. (2016). *The news media: What everyone needs to know.* Oxford University Press.

Anderson, C.W. (2011). Deliberative, agonistic, and algorithmic audiences: Journalism's vision of its public in an age of audience transparency. *International Journal of Communication*, 5, 19.

Anderson, C.W. (2013). What aggregators do: Towards a networked concept of journalistic expertise in the digital age. *Journalism: Theory, Practice & Criticism*, 14(8), 1008–1023.

Andreessen, M. (2011). Why software is eating the world. *Wall Street Journal*, 20(2011), C2.

Andrejevic, M. (2014). Big data, big questions. The big data divide. *International Journal of Communication*, 8, 17.

Andrew, D. (2019). Programmatic trading: the future of audience economics. *Communication Research and Practice*, 5, 73–87.

Angrist, J. D., & Pischke, J. S. (2014). *Mastering metrics: The path from cause to effect*. Princeton University Press.

Annual Iowa State Report. Vol. 12, 1889.

Atkinson, R., & Flint, J. (2001). Accessing hidden and hard-to-reach populations: Snowball research strategies. *Social research update*, 33(1), 1–4.

Attewell, P. (1992). Technology diffusion and organizational learning: The case of business computing. *Organization science*, 3(1), 1–19.

Axel Springer (2019, March 6). *Axel Springer acquires paid content technology company CeleraOne.* https://www.axelspringer.com/en/ax-press-release/axel-springer-acquires-paid-content-technology-company-celeraone.

Baack, S. (2015). Datafication and empowerment: How the open data movement re-articulates notions of democracy, participation, and journalism. *Big Data & Society*, 2(2), 2053951715594634.

Baack, S. (2016). What big data leaks tell us about the future of journalism–and its past. *Internet policy review*, 12(23), 9–17.

Baack, S. (2018). Practically engaged: The entanglements between data journalism and civic tech. *Digital Journalism*, 6(6), 673–692.

Baesens, B., Roesch, D., & Scheule, H. (2016). *Credit risk analytics: Measurement techniques, applications, and examples in SAS.*

Balazka, D., & Rodighiero, D. (2020). Big data and the little big bang: an epistemological (R)evolution. *Frontiers in big Data*, 3, 31.

Barnett, P. (2015). If what gets measured gets managed, measuring the wrong thing matters. *Corporate Finance Review*, 19(4), 5.

Baxter, P., & Jack, S. (2008). Qualitative case study methodology: Study design and implementation for novice researchers. *The qualitative report*, 13(4), 544–559.

Beam, R. A. (1995). How Newspapers Use Readership Research. *Newspaper Research Journal*, 16(2), 28–38.

Bechmann, A., Bilgrav-Nielsen, K., & Korsgaard Jensen, A. L. (2016). Data as a revenue model: Sharewall as a payment method and editorial algorithm in the news industry. *Nordicom information*, 38(1), 76–82.

Beer, D. (2016). *Metric Power*. Palgrave Macmillan UK.

Beer, D. (2019). *The Data Gaze: Capitalism, Power and Perception*. SAGE Publications Ltd.

Belair-Gagnon, V., & Holton, A. E. (2018). Boundary Work, Interloper Media, and Analytics in Newsrooms: An Analysis of the roles of Web Analytics Companies in News Production. *Digital Journalism*, 6(4), 492–508.

Bell, E. & Owen, T. (2017, March 29). The Platform Press: How Silicon Valley reengineered journalism. *Columbia Journalism Review*, https://www.cjr.org/tow_center_reports/platform-press-how-silicon-valley-reengineered-journalism.php

Benbasat, I., Goldstein, D. K., & Mead, M. (1987). The case research strategy in studies of information systems. *MIS quarterly*, 369–386.

Benn, S., Edwards, M., & Angus-Leppan, T. (2013). Organizational learning and the sustainability community of practice: The role of boundary objects. *Organization & Environment*, 26(2), 184–202.

Bergermann, U., Hanke, C. (2017). Boundary Objects, Boundary Media. Von Grenzobjekten und Medien bei Susan Leigh Star und James R. Griesemer. In: Gießmann, S., Taha, N. (Eds). *Grenzobjekte und Medienforschung.* 117–130.

Billsus, D. (2000). User Modeling for Adaptive News Access. *User Modeling and User-Adapted Interaction*, 10(2/3), 147–180.

Blanchett, N. (2021). Participative gatekeeping: The intersection of news, audience data, newsworkers, and economics. *Digital Journalism*, 9(6), 773–791.

Boczkowski, P. J. (2005). *Digitizing the news: Innovation in online newspapers.* MIT Press.

Bodó, B. (2021). Selling News to Audiences – A Qualitative Inquiry into the Emerging Logics of Algorithmic News Personalization in European Quality News Media. *Digital Journalism*, 7(8), 1–22.

Bodó, B., Helberger, N., Eskens, S., & Möller, J. (2019). Interested in diversity: The role of user attitudes, algorithmic feedback loops, and policy in news personalization. *Digital Journalism*, 7(2), 206–229.

Bogner, A., & Menz, W. (2009). The theory-generating expert interview: epistemological interest, forms of knowledge, interaction. In: Bogner, A., Littig, B., & Menz, W. (Eds.). *Interviewing Experts.* 43–80.

Bortolini, R. F., Nogueira Cortimiglia, M., Danilevicz, A. D. M. F., & Ghezzi, A. (2021). Lean Startup: a comprehensive historical review. *Management Decision*, 59(8), 1765–1783.

Boyd, D., & Crawford, K. (2011). Six provocations for big data. In: *A decade in internet time: Symposium on the dynamics of the internet and society.*

Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5), 662–679.

Boyles, J. L. (2020). Deciphering Code: How Newsroom Developers Communicate Journalistic Labor. *Journalism Studies*, 21(3), 336–351.

Brandtzaeg, P. B., Lüders, M., Spangenberg, J., Rath-Wiggins, L., & Følstad, A. (2016). Emerging journalistic verification practices concerning social media. *Journalism Practice*, 10(3), 323–342.

Brandtzaeg, P. B., & Følstad, A. (2017). Trust and distrust in online fact-checking services. *Communications of the ACM*, 60(9), 65–71.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2), 77–101.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1–7), 107–117.

Brinkmann, S. & Kvale, S. (2018). *Doing Interviews*. Sage.

Bro, P., & Wallberg, F. (2015). Gatekeeping in a digital era: Principles, practices and technological platforms. *Journalism Practice*, 9(1), 92–105.

Broby, D. (2022). The use of predictive analytics in finance. *The Journal of Finance and Data Science*, 8, 145–161.

Bryant, L. R. (2011). *The democracy of objects.* Open Humanities Press.

Bryman, A. (2016). *Social research methods.* Oxford University Press.

Buckland, M. K. (1991). Information as thing. *Journal of the American Society for information science*, 42(5), 351–360.

Bunce, M. (2019). Management and resistance in the digital newsroom. *Journalism*, 20(7), 890–905.

Buschow, C., & Wellbrock, C. M. (2020). *Money for Nothing and Content for Free?.* Nomos.

Bucher, T. (2018). *If... then:* Algorithmic power and politics. Oxford University Press.

Busch, O. (2014). *Programmatic advertising.* The successful transformation to automated, data-driven marketing in real-time.

Cabra, M., & Kissane, E. (2016). Wrangling 2.6 TB of data: The people and the technology behind the Panama Papers. *International Consortium of Investigative Journalists.*

Caplan, R., & Boyd, D. (2018). Isomorphism through algorithms: Institutional dependencies in the case of Facebook. *Big Data & Society*, 5(1), 205395171875725.

(2008) Cambridge advanced learner's dictionary. Cambridge University Press.

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and program planning*, 2(1), 67–90.

Capurro, R. (2009). Past, present, and future of the concept of information. *TripleC: communication, capitalism & critique. open access journal for a global sustainable information society*, 7(2), 125–141.

Carlson, M. (2015). The Robotic Reporter: Automated journalism and the redefinition of labor, compositional forms, and journalistic authority. *Digital Journalism*, 3(3), 416–431.

Carlson, M., & Lewis, S. C. (Eds.). (2015). *Boundaries of journalism: Professionalism, practices and participation.* Routledge.

Carmi, E., Yates, S. J., Lockley, E., & Pawluczuk, A. (2020). Data citizenship: Rethinking data literacy in the age of disinformation, misinformation, and malinformation. *Internet Policy Review*, 9(2), 1–22

Carroll, E. (2020). *News as Surveillance.* Georgetown University Law Center. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3516731

Caswell, D. (2021). Structured journalism and the semantic units of news. In: Thurman, N., Lewis, S. C., & Kunert, J. (Eds.). *Algorithms, automation, and news: New directions in the study of computation and journalism.* 155–177, Routledge.

Cavaye, A. L. (1996). Case study research: a multi-faceted research approach for IS. *Information systems journal*, 6(3), 227–242.

Chakraborty, A. (2018). On Designing Content Recommender Systems for Online News Media. *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems – CHI '18.*

Chakraborty, A., Luqman, M., Satapathy, S., & Ganguly, N. (2018). Editorial Algorithms. *Companion of the The Web Conference 2018 on The Web Conference 2018 – WWW '18.*

Chandler, A. D. (1978). *The visible hand.* The Managerial Revolution in American Business (Cambridge, MA, 1977).

Cherubini, F., & Nielsen, R. K. (2016). Editorial analytics: How news media are developing and using audience data and metrics. *SSRN 2739328.*

Chesney, R., & Citron, D. (2019). Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Affairs*, 98, 147.

Chipman, H. A., & Joseph, V. R. (2016). A Conversation with Jeff Wu. Statistical Science, 31(4), 624–636.

Chomsky, N. (1966). Explanatory models in linguistics. In *Studies in Logic and the Foundations of Mathematics* (Vol. 44, pp. 528–550). Elsevier.

Christmann, G. B. (2009). Expert Interviews on the Telephone: A Difficult Undertaking. In: Bogner, A., Littig, B., Menz, W. (Eds.). *Interviewing Experts.* 157–183.

Chyi, H. I., & Tenenboim, O. (2016). Reality check: Multiplatform newspaper readership in the United States, 2007–2015. *Journalism Practice*, 11(7), 798–819.

Coddington, M. (2014). Clarifying Journalism's Quantitative Turn. *Digital Journalism*, 3(3), 331–348.

Coddington, M. (2015). The wall becomes a curtain. *Boundaries of journalism*, 67–82.

Coddington, M. (2020). Gathering evidence of evidence: News aggregation as an epistemological practice. *Journalism*, 21(3), 365–380.

Cornia, A., Sehl, A., & Nielsen, R. K. (2020). 'We no longer live in a time of separation': A comparative analysis of how editorial and commercial integration became a norm. *Journalism*, 21(2), 172–190.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

Costera Meijer, I., & Groot Kormelink, T. (2015). Checking, Sharing, Clicking and Linking: Changing patterns of news use between 2004 and 2014. *Digital Journalism*, 3(5), 664–679.

Cottle, S., & Ashton, M. (1999). From BBC newsroom to BBC newscentre: On changing technology and journalist practices. *Convergence*, 5(3), 22–43.

Couldry, N. (2012). *Media, Society, World: Social Theory and Digital Media Practice.*

Couldry, N. (2020). Recovering critique in an age of datafication. *New Media & Society*, 22(7), 1135–1151.

Couldry, N., & Mejias, U. A. (2019). Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject. *Television & New Media*, 20(4), 336–349.

Coulmas, F. (1989). *The writing systems of the world.*

Crawford K. (2014). The Test We Can—and Should—Run on Facebook. How to reclaim power in the era of perpetual experiment engines. *The Atlantic*. July 2, 2014.

Creswell, J. W., & Poth, C. N. (2018). *Qualitative inquiry and research design: Choosing among five approaches.* Sage.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches.* Sage.

Cukier, K. & Mayer-Schönberger, V. (2013). The Rise of Big Data: How It's Changing the Way We Think About the World. *Foreign Affairs*, 92(3), 28–40.

Cushion, S., Lewis, J., & Callaghan, R. (2017). Data journalism, impartiality and statistical claims: Towards more independent scrutiny in news reporting. *Journalism Practice*, 11(10), 1198–1215.

Czarniawska-Joerges, B. (1997). *Narrating the organization: Dramas of institutional identity*. University of Chicago Press.

D'Ignazio, C., & Klein, L. F. (2023). *Data feminism.* MIT Press.

Dalton, C., & Thatcher, J. (2014). What does a critical data studies look like, and why do we care? Seven points for a critical approach to 'big data'. *Society and Space*, 29.

Dalton, C. M., Taylor, L., & Thatcher, J. (2016). Critical data studies: A dialog on data and space. *Big Data & Society*, 3(1), 2053951716648346.

Darcy, O. (2023, June 22). Vice Media, once worth billions, set to be acquired out of bankruptcy by its creditors for $350 million. *CNN Business*, https://edition.cnn.com/2023/06/22/media/vice-acquired-bankruptcy.

Davenport, T. H., & Patil, D. J. (2012). Data scientist. H*arvard business review*, 90(5), 70–76.

Davenport, T. H., & Patil, D. J. (2022). Is data scientist still the sexiest job of the 21st century. *Harvard Business Review*, 15.

Dencik, L., & Kaun, A. (2020). Datafication and the welfare state. *Global Perspectives*, 1(1), 12912.

Desai, A., Nouvellet, P., Bhatia, S., Cori, A., & Lassmann, B. (2021). Data journalism and the COVID-19 pandemic: opportunities and challenges. *The Lancet Digital Health*, 3(10), 619–621.

Deuze, M. (2012). *Media life*. Polity.

Deuze, M., & Witschge, T. (2018). Beyond journalism: Theorizing the transformation of journalism. *Journalism*, 19(2), 165–181.

Devriendt, F., Berrevoets, J., & Verbeke, W. (2021). Why you should stop predicting customer churn and start using uplift models. *Information Sciences*, 548, 497–515. https://doi.org/10.1016/j.ins.2019.12.075

DeMarco, T. (1986). *Controlling software projects:* Management, measurement, and estimates. Prentice Hall PTR.

Diaconis, P. (2006). *Theories of data analysis: From magical thinking through classical statistics.* Exploring data tables, trends, and shapes, 1–36.

Diakopoulos, N. (2014). Algorithmic Accountability. *Digital Journalism*, 3(3), 398–415.

Diakopoulos, N. (2019). *Automating the News:* How Algorithms Are Rewriting the Media. Harvard University Press.

Dierickx, L. (2021). News automation, materialities, and the remix of an editorial process. *Journalism*, 146488492110238.

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, 147–160.

Dourish, P., & Gómez Cruz, E. (2018). Datafication and data fiction: Narrating data and narrating with data. *Big Data & Society*, 5(2), 205395171878408.

Dörr, K. N. (2015). Mapping the field of Algorithmic Journalism. *Digital Journalism*, 4(6), 700–722.

Drucker, P. (1966). The effective manager. *Organization*, 1, 2.

Drucker, P. (1996). *Landmarks of Tomorrow: a report on the new post-modern world.* Routledge.

Dubé, L., & Paré, G. (2003). Rigor in information systems positivist case research: current practices, trends, and recommendations. *MIS quarterly*, 597–636.

Duffy, A., Tandoc, E. C., & Ling, R. (2018). Frankenstein journalism. *Information, Communication & Society*, 21(10), 1354–1368.

Durand, J. D. (1967). The modern expansion of world population. *Proceedings of the American Philosophical Society,* 111(3), 136–159.

Edmondson, A. C., & McManus, S. E. (2007). Methodological fit in management field research. Academy of management review, 32(4), 1246–1264.

Eckerson, W. W. (2007). Predictive analytics. Extending the Value of Your Data Warehousing Investment. *TDWI Best Practices Report*, 1, 1–36.

Eckerson, W. W. (2010). *Performance dashboards: Measuring, monitoring, and managing your business.*

Eisenhardt, K. M. (1989). Building theories from case study research. Academy of management review, 14(4), 532–550.

Eisenhardt, K. M., & Graebner, M. E. (2007). Theory building from cases: Opportunities and challenges. *Academy of management journal*, 50(1), 25–32.

Ekberg, A. S. K. (2018). *The Role of Organizational Integrity in Responses to Pressures:* A Case Study of Australian Newspapers [PhD, Queensland University of Technology].

Ekdale, B., Singer, J. B., Tully, M., & Harmsen, S. (2015). Making Change: Diffusion of Technological, Relational, and Cultural Innovation in the Newsroom. *Journalism & Mass Communication Quarterly*, 92(4), 938–958.

Ekström, M., Ramsälv, A., & Westlund, O. (2022). Data-driven news work culture: Reconciling tensions in epistemic values and practices of news journalism. *Journalism*, 23(4), 755–772.

Eldridge II, S. A. (2017). *Online journalism from the periphery:* Interloper media and the journalistic field. Routledge.

Eldridge II, S. A. (2019). Where do we draw the line? Interlopers, (ant)agonists, and an unbounded journalistic field. *Media and Communication*, 7(4), 8–18.

EU Commission (2016). 2016/679 of the European Parliament and of the Council. *Official Journal of the European Union*, 679, 2016, https://eur-lex.europa.eu/eli/reg/2016/679/oj.

EU Commission (2017). Proposal for a Regulation on Privacy and Electronic Communications. *COM(2017) 10 final,* 10 Januar 2017.

EU Commision (2020). *Commission proposes measures to boost data sharing and support European data spaces,* 25 November 2020. https://ec.europa.eu/commission/presscorner/detail/en/ip_20_2102

EU Commission (2022). Proposal for a Regulation of the European Parliament and of the Council on harmonized rules on fair access to and use of data (Data Act), *COM(2022) 68 final*, 23 February 2022.

Evens, T., & Van Damme, K. (2016). Consumers' willingness to share personal data: Implications for newspapers' business models. *International Journal on Media Management*, 18(1), 25–41.

Fails, J. A., & Olsen Jr, D. R. (2003). Interactive machine learning. In: *Proceedings of the 8th international conference on Intelligent user interfaces*, 39–45.

Feinberg, M., Sutherland, W., Nelson, S. B., Jarrahi, M. H., & Rajasekar, A. (2020). The new reality of reproducibility: The role of data work in scientific research. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1), 1–22.

Filipović, A. (2015). Die Datafizierung der Welt. Eine ethische Vermessung des digitalen Wandels. *Communicatio Socialis (ComSoc)*, 48(1), 6–15.

Fiske, A., Prainsack, B., & Buyx, A. (2019). Data Work: Meaning-Making in the Era of Data-Rich Medicine. *Journal of Medical Internet Research*, 21(7), e11672.

Flensburg, S., & Lomborg, S. (2021). Datafication research: Mapping the field for a future agenda. *New Media & Society*, 146144482110466.

Fletcher, R., & Nielsen, R. K. (2018). Automated serendipity: The effect of using search engines on news repertoire balance and diversity. *Digital Journalism*, 6(8), 976–989.

Flew, T., Spurgeon, C., Daniel, A., & Swift, A. (2012). The Promise of Computational Journalism. *Journalism Practice*, 6(2), 157–171.

Floridi, L. (2010). *Information:* A very short introduction. Oxford University Press.

Flyvbjerg, B. (2006). Five Misunderstandings About Case-Study Research. *Qualitative Inquiry*, 12(2), 219–245.

Flyvbjerg, B. (2011). Case study. In: Denzin, N. K., Lincoln, Y. S. (Eds.): *The Sage Handbook of Qualitative Research*, 4, 301–316.

Flyverbom, M., & Murray, J. (2018). Datastructuring—Organizing and curating digital traces into action. *Big Data & Society*, 5(2), 205395171879911.

Foucault, M. (1980a). *Language, counter-memory, practice: Selected essays and interviews.* Cornell University Press.

Foucault, M. (1980b). *Power/knowledge* (C. Gordon, trans.). Pantheon.

Foucault, M. (2003). *The Birth of the Clinic.* Routledge. Original work published 1963.

Fürnkranz, J. (2011). Decision Tree. In: Sammut, C., Webb, G.I. (Eds.). *Encyclopedia of Machine Learning.* Springer, Boston, MA.

Franklin, B., & Eldridge II, S. (Eds.). (2016). *The Routledge companion to digital journalism studies.* Taylor & Francis.

Froschauer, U., & Lueger, M. (1992). *Das qualitative Interview:* Zur Analyse sozialer Systeme. WUV, Univ.-Verlag.

Fuchs, C. (2010). Grounding Critical Communication Studies: An Inquiry into the Communication Theory of Karl Marx. *Journal of Communication Inquiry*, 34(1), 15–41.

Fuchs, C., & Mosco, V. (2016). *Marx in the Age of Digital Capitalism.* Brill. https://doi.org/10.1163/9789004291393

Fuller, M., & Goffey, A. (2012). *Evil Media.* The MIT Press. https://doi.org/10.7551/mitpress/8696.001.0001

Galton, F. (1904). Eugenics: Its definition, scope, and aims. *American Journal of Sociology*, 10(1), 1-25.

Gade, P. J., & Lowrey, W. (2011). Reshaping the journalistic culture. *Changing the news:* The forces shaping journalism in uncertain times, 22–42.

Gara, A., Hughes, J. & Paolo Mancini, D. (2023). Nasdaq to buy Adenza for $10.5bn in US exchange operator's biggest deal, *Financial Times*, https://ft.com/content/bf188909-6577-404b-89c2-79a8261e2e0e.

Geertz, C. (1973). *The interpretation of cultures* (Vol. 5019).

Gellert, R. (2022). Comparing definitions of data and information in data protection law and machine learning: A useful way forward to meaningfully regulate algorithms? *Regulation & governance*, 16(1), 156–176.

George, A. L., Bennett, A. (2005). *Case Studies and Theory Development in the Social Sciences.* The MIT Press.

Gibbert, M., Ruigrok, W., & Wicki, B. (2008). What passes as a rigorous case study?. *Strategic management journal*, 29(13), 1465–1474.

Gieryn, T. F. (1983). Boundary-work and the demarcation of science from non-science: Strains and interests in professional ideologies of scientists. *American Sociological Review*, 781–795

Gitelman, L. (Ed.). (2013). *Raw data is an oxymoron.* The MIT Press.

Gladwell, M. (2008). *Outliers: The story of success.* Little, Brown.

Glaser, B., & Strauss, A. (1999). *Discovery of grounded theory: Strategies for qualitative research.* Routledge.

Gläser, J., & Laudel, G. (2010). *Experteninterviews und qualitative Inhaltsanalyse.* Springer-Verlag.

Graus, D., & Dumtrache, A. (2020). *Beyond Optimizing for Clicks:* Incorporating Editorial Values in News Recommendation.

Gray, J., Chambers, L., & Bounegru, L. (2012). *The data journalism handbook: How journalists can use data to improve the news.* O'Reilly.

Greenwood, R., Raynard, M., Kodeih, F., Micelotta, E. R., & Lounsbury, M. (2011). Institutional complexity and organizational responses. Academy of Management annals, 5(1), 317-371.

Gröger, C. (2021). There is no AI without data. *Communications of the ACM*, 64(11), 98–108.

Guaglione, S. (2024, February 8). The New York Times expects ad revenue to continue to decline in 2024. *Digiday*, https://digiday.com/media/the-new-york-times-expects-ad-revenue-to-continue-to-decline-in-2024/.

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods*, 18(1), 59–82.

Guevara, M.W. (2019, March 25). How Artificial Intelligence Can Help Us Crack More Panama Papers Stories. *International Consortium of Investigative Journalists*, https://www.icij.org/inside-icij/2019/03/how-artificial-intelligence-can-help-us-crack-more-panama-papers-stories/.

Hacking, I. (1992). Making Up People. In: Stein, E. (Ed.). *Forms of desire.* Sexual orientation and the social constructionist controversy, 69–88.

Hacking, I. (1999). *The social construction of what?* Harvard University Press.

Hacking, I. (2002). Historical ontology. In: In the Scope of Logic, Methodology and Philosophy of Science. *Volume Two of the 11th*

International Congress of Logic, Methodology and Philosophy of Science, Cracow, August 1999, 583–600. Springer Netherlands.

Hacking, I. (2007, April). Kinds of people: Moving targets. In: *Proceedings British Academy* (Vol. 151, 285). Oxford University Press Inc..

Hahn, D. (1996). Strategisches und operatives Controlling. In: *PuK — Controllingkonzepte*. Gabler Verlag, Wiesbaden.

Haim, M., Graefe, A., & Brosius, H. B. (2018). Burst of the filter bubble? Effects of personalization on the diversity of Google News. *Digital Journalism*, 6(3), 330–343.

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE intelligent systems*, 24(2), 8–12.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining concepts and techniques,* third edition.

Hanitzsch, T. (2019). Journalism studies still needs to fix Western bias. *Journalism*, 20(1), 214–217.

Hanusch, F. (2017). Web analytics and the functional differentiation of journalism cultures: Individual, organizational and platform-specific influences on newswork. *Information, Communication & Society*, 20(10), 1571–1586.

Harambam, J., Bountouridis, D., Makhortykh, M., & van Hoboken, J. (2019). Designing for the better by taking users into account: A qualitative evaluation of user control mechanisms in (news) recommender systems. *Proceedings of the 13th ACM Conference on Recommender Systems – RecSys '19*, 69–77.

Harambam, J., Helberger, N., & van Hoboken, J. (2018). Democratizing algorithmic news recommenders: How to materialize voice in a technologically saturated media ecosystem. *Philosophical Transactions of the Royal Society A:* Mathematical, Physical and Engineering Sciences, 376(2133), 20180088.

Hartmann, M. (2018). *Die Abgehobenen: wie die Eliten die Demokratie gefährden*. Campus Verlag.

Heft, A. (2019). The Panama Papers investigation and the scope and boundaries of its networked publics: Cross-border journalistic collaboration driving transnationally networked public spheres. *Journal of Applied Journalism & Media Studies*, 8(2), 191–209.

Helberger, N. (2019). On the Democratic Role of News Recommenders. *Digital Journalism*, 1–20.

Helberger, N., Karppinen, K., & D'Acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, 21(2), 191–207.

Hendrickx, J., & Picone, I. (2020). Innovation Beyond the Buzzwords: The Rocky Road Towardsa Digital First-based Newsroom. *Journalism Studies*, 21(14), 2025–2041. https://doi.org/10.1080/1461670X.2020.1809494

Hepp, A. (2019). *Deep Mediatization.* 1, Routledge.

Heravi, B. (2019). 3WS of Data Journalism Education: What, where and who?. *Journalism Practice*, 13(3), 349–366.

Heravi, B., Cassidy, K., Davis, E., & Harrower, N. (2022). Preserving data journalism: A systematic literature review. *Journalism Practice*, 16(10), 2083–2105.

Herman, E. S., Chomsky, N. (2010). *Manufacturing consent: The political economy of the mass media.* Random House.

Hirsch, L. & Mullin, B. (2023, June 22). Fortress Investment Group Set to Acquire Vice Out of Bankruptcy. *New York Times*, https://www.nytimes.com/2023/06/22/business/fortress-vice-bankruptcy.html.

Hockenhull, M., & Cohn, M. L. (2021). Speculative data work & dashboards: designing alternative data visions. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3), 1–31.

Holton, A. E., & Belair-Gagnon, V. (2018). Strangers to the game? Interlopers, intralopers, and shifting news production. *Media and communication*, 6(4), 70–78.

Holton, A. E., Coddington, M., & Gil de Zúñiga, H. (2013). Whose news? Whose values? Citizen journalism and journalistic values through the lens of content creators and consumers. *Journalism Practice*, 7(6), 720–737.

Hoppit, J. (1987). Understanding the Industrial Revolution. *The Historical Journal*, 30(1), 211–224.

Howard, A.B. (2014). *The art and science of data-driven journalism.*

Hughes, A. (2010). Innovation Policy as Cargo Cult: Myth and Reality in Knowledge-Led Productivity Growth. In: López-Claros, A. (Eds). *The Innovation for Development Report 2009–2010*. Strengthening Innovation for the Prosperity of Nations.

Hussey, N. E., Kessel, S. T., Aarestrup, K., Cooke, S. J., Cowley, P. D., Fisk, A. T., Harcourt, R.G., Holland, K.N., Iverson, S.J., Kocik, J.F., Mills Flemming, J.E. & Whoriskey, F. G. (2015). Aquatic animal telemetry: a panoramic window into the underwater world. Science, 348(6240), 1255642.

Huvila, I., Anderson, T. D., Jansen, E. H., McKenzie, P., & Worrall, A. (2017). Boundary objects in information science. *Journal of the Association for Information Science and Technology*, 68(8), 1807–1822.

Iliadis, A., & Russo, F. (2016). Critical data studies: An introduction. *Big Data & Society*, 3(2), 2053951716674238.

Ivanova, I. (2023, August 17). What if OpenAI trained ChatGPT with illegal data scraping? The New York Times is reportedly considering suing to put that to the test. *Fortune*, https://fortune.com/2023/08/17/openai-new-york-times-lawsuit-illegal-scraping.

Jadhav, R. J., & Pawar, U. T. (2011). Churn prediction in telecommunication using data mining technology. *International Journal of Advanced Computer Science and Applications*, 2(2).

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information systems management*, 29(4), 258–268.

Johnson, J. M., & Rowlands, T. (2012). The Interpersonal Dynamics of In-Depth Interviewing. In: Gubrium, J. F., Holstein, J. A., Marvasti, A. B., & McKinney, K. D. (Eds.). *The SAGE Handbook of Interview Research.* The Complexity of the Craft, 99–114.

Kalleberg, A. L. (Ed.). (1996). *Organizations in America:* Analysing their structures and human resource practices.

Karatzoglou, A., & Hidasi, B. (2017). Deep Learning for Recommender Systems. *Proceedings of the Eleventh ACM Conference on Recommender Systems—RecSys '17.* https://doi.org/10.1145/3109859.3109933

Karim Schapals, A. (2022). *Peripheral actors in journalism:* Deviating from the norm? Routledge.

Karimi, M., Jannach, D., & Jugovac, M. (2018). News recommender systems – Survey and roads ahead. *Information Processing & Management*, 54(6), 1203–1227.

Kaushik, A. (2009). *Web analytics 2.0: The art of online accountability and science of customer centricity.* John Wiley & Sons.

Keen, W. W., Blake, C. H. (1927). Datum and Data. *Science*, 66(1696), 1927, 15–15.

Keller, S., & Price, C. (2011). *Beyond performance.* How Great Organizations Build Ultimate Competitive Advantage.

Kelvin, W. T. B. (1889). Electrical Units of Measurement. *Popular lectures and addresses* (Vol. 1). Macmillan and Company.

Kennedy, H., Oman, S., Taylor, M., Bates, J., & Steedman, R. (2021). *Public understanding and perceptions of data practices:* A review of existing research.

Kenessey, Z. (1987). The primary, secondary, tertiary and quaternary sectors of the economy. *Review of income and wealth*, 33(4), 359–385.

Kitchin, R. (2022). *The Data Revolution:* A critical analysis of big data, open data and data infrastructures.

Kitchin, R., & Lauriault, T. (2014). *Towards Critical data studies:* Charting and unpacking data assemblages and their work.

Kline, R.R. (2015). *The Cybernetics Moment.* Or Why We Call Our Age the Information Age, John Hopkins University Press.

Koskinen, K. U., & Mäkinen, S. (2009). Role of boundary objects in negotiations of project contracts. *International Journal of Project Management*, 27(1), 31–38.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1), 3–24.

Kosterich, A. (2020). Managing news nerds: Strategizing about institutional change in the news industry. *Journal of Media Business Studies*, 17(1), 51–68.

Kuckartz, U. (2019). *Qualitative text analysis: A systematic approach. Compendium for early career researchers in mathematics education.* 181–197.

Kumar, V., & Garg, M. L. (2018). Predictive analytics: a review of trends and techniques. *International Journal of Computer Applications*, 182(1), 31–37.

Kumar, V., & Reinartz, W. (2018). *Customer relationship management.*

Kvale, S., & Brinkmann, S. (2009). Interviews: Learning the craft of qualitative research interviewing. Sage.

Lamont, M., & Molnár, V. (2002). The study of boundaries in the social sciences. *Annual Review of Sociology*, 28(1), 167–195.

Larose, D. T. (2015). *Data mining and predictive analytics.*

Latour, B., & Woolgar, S. (1986). *Laboratory life:* The construction of scientific facts. Princeton University Press.

Latour, B. (1992). Where are the missing masses? The sociology of a few mundane artifacts. *Shaping technology/building society:* Studies in sociotechnical change, 1, 225–258.

Latour, B. (1999). *Pandora's Hope:* Essays on the Reality of Science Studies. 1st edition. Cambridge, Mass: Harvard University Press.

Lauer, T. (2020). *Change management: fundamentals and success factors.*

Lawless, J. F. (2011). *Statistical models and methods for lifetime data.* John Wiley & Sons.

Lee, C. P. (2007). Boundary negotiating artifacts: Unbinding the routine of boundary objects and embracing chaos in collaborative work. *Computer Supported Cooperative Work (CSCW)*, 16, 307–339.

Leonard, T. C., Thaler, R., Sunstein, C. (2008). *Nudge: Improving decisions about health, wealth, and happiness.* Yale University Press.

Lewis, S. C., & Usher, N. (2013). Open source and journalism: Toward new frameworks for imagining news innovation. *Media, Culture & Society*, 35(5), 602–619.

Lewis, S. C., & Usher, N. (2016). Trading zones, boundary objects, and the pursuit of news innovation: A case study of journalists and programmers. *Convergence: The International Journal of Research into New Media Technologies*, 22(5), 543–560.

Lewis, S. C., Guzman, A. L., & Schmidt, T. R. (2019). Automation, Journalism, and Human–Machine Communication: Rethinking Roles and Relationships of Humans and Machines in News. *Digital Journalism*, 7(4), 409–427.

Li, X., Li, X., Li, Z., Ma, M., Wang, J., Xiao, Q., Liu, Q., Che, T., Chen, E., Yan, G., Hu, Z., Zhang, L., Chu, R., Su, P., Liu, Q., Liu, S., Wang, J., Niu, Z., Chen, Y., Jin, R., Wang, W., Ran, Y., Xin, X. & Ren, H. (2009). Watershed allied telemetry experimental research. *Journal of Geophysical Research: Atmospheres*, 114(D22).

Lindberg, T. (2023). *Nordic News Media in Global Competition.*

Lindlof, T. R., & Taylor, B. C. (2017). *Qualitative communication research methods.* Sage.

Ling, C. X., & Li, C. (1998). Data mining for direct marketing: Problems and solutions. *The Fourth International Conference on Knowledge Discovery and Data Mining—KDD-98* (Vol. 98, pp. 73-79).

Link, J. (1982): Die methodologischen, informationswirtschaftlichen und führungspolitischen Aspekte des Controlling. *Zeitschrift für Betriebswirtschaft*, 52, 261–280.

Lischka, J. A., Schaetz, N., & Oltersdorf, A.-L. (2022). Editorial Technologists as Engineers of Journalism's Future: Exploring the Professional Community of Computational Journalism. *Digital Journalism*, 1–19.

Liu, J., Dolan, P., & Pedersen, E. R. (2010, February). Personalized news recommendation based on click behavior. In: *Proceedings of the 15th international conference on Intelligent user interfaces*, 31–40.

Logan, R. K. (2012). What is information?: Why is it relativistic and what is its relationship to materiality, meaning and organization. *Information*, 3(1), 68–91.

Loosen, W., Ahva, L., Reimer, J., Solbach, P., Deuze, M., & Matzat, L. (2022). 'X Journalism'. Exploring journalism's diverse meanings through the names we give it. *Journalism*, *23*(1), 39–58.

Loosen, W., Pörksen, B., & Scholl, A. (2008). Paradoxien des Journalismus: Einführung und Begriffsklärung. *Paradoxien des Journalismus: Theorie—Empirie—Praxis*, Festschrift für Siegfried Weischenberg, 17–33.

Loosen, W., Reimer, J., & De Silva-Schmidt, F. (2020). Data-driven reporting: An on-going (r)evolution? An analysis of projects nominated for the Data Journalism Awards 2013–2016. *Journalism*, 21(9), 1246–1263.

Loosen, W., Solbach, P. (2020). Künstliche Intelligenz im Journalismus? Was bedeutet Automatisierung für journalistisches Arbeiten? In: Köhler, T. (Ed.). *Fake News, Framing, Fact-Checking: Nachrichten im digitalen Zeitalter*. Transcript, 177–204.

Loukissas, Y. A. (2019). *All data are local: Thinking critically in a data-driven society*. MIT Press.

Lowrey, W., & Hou, J. (2021). All forest, no trees? Data journalism and the construction of abstract categories. *Journalism*, 22(1), 35–51.

Lucey, T. (2004). *Management information systems.* Int. Thomson Business Press.

Maares, P., & Hanusch, F. (2020). Exploring the boundaries of journalism: Instagram micro-bloggers in the twilight zone of lifestyle journalism. *Journalism*, 21(2), 262–278.

MacCormack, A., Baldwin, C., & Rusnak, J. (2012). Exploring the duality between product and organizational architectures: A test of the "mirroring" hypothesis. *Research Policy*, 41(8), 1309–1324.

Machlup, F. (1983). *The study of information:* Interdisciplinary messages.
Mackenzie, A. & Mills, R. et al. (2015). Digital sociology in the field of devices. *Routledge International Handbook of the Sociology of Art and Culture*, 367–382.

Mager, A., & Katzenbach, C. (2021). Future imaginaries in the making and governing of digital technology: Multiple, contested, commodified. *New Media & Society*, 23(2), 223–236.

Malterud, K., Siersma, V. D., & Guassora, A. D. (2016). Sample size in qualitative interview studies: guided by information power. *Qualitative health research*, 26(13), 1753–1760.

March, J. G., & Olsen, J. P. (2011). *The logic of appropriateness.*
Marres, N. (2017). *Digital sociology:* The reinvention of social research. John Wiley & Sons.

Marshall, C., & Rossman, G. B. (2016). *Designing qualitative research.* 6ed, Sage.

Marshall, B., Cardon, P., Poddar, A., & Fontenot, R. (2013). Does sample size matter in qualitative research?: A review of qualitative interviews in IS research. *Journal of computer information systems*, 54(1), 11–22.

Marx, K. (2005). Grundrisse: Foundations of the Critique of Political Economy.

Mason, M. (2010). Sample size and saturation in PhD studies using qualitative interviews. *Forum qualitative Sozialforschung*, 11(3).

Matthes, J., & Kohring, M. (2008). The content analysis of media frames: Toward improving reliability and validity. *Journal of Communication*, 58(2), 258–279.

McLoughlin, I., Rose, H., & Clark, J. (1985). Managing the introduction of new technology. *Omega*, 13(4), 251–262.

Merton, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science*, 159(3810), 56–63.

Meuser, M., & Nagel, U. (2009). The expert interview and changes in knowledge production. In: Bogner, A., Littig, B., Menz, W. (Eds.). *Interviewing experts* (17–42). Palgrave Macmillan UK.

Mey, G., & Mruck, K. (2011). *Grounded theory reader (2)*. VS Verlag für Sozialwissenschaften.

Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology*, 83(2), 340–363.

Miceli, M., & Posada, J. (2022). The Data-Production Dispositif. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), 1–37.

Mieg, H. A., & Näf, M. (2005). *Experteninterviews.* Institut für Mensch-Umwelt-Systeme (HES), ETH Zürich.

Mitchell, T. M. (1997). Machine learning.

Mitkov, R. (2014). *Anaphora resolution.* Routledge.

Møller, N. H., Bossen, C., Pine, K. H., Nielsen, T. R., & Neff, G. (2020). Who does the work of data? *Interactions*, 27(3), 52–55.

Moran, R. E., & Shaikh, S. J. (2022). Robots in the news and newsrooms: Unpacking meta-journalistic discourse on the use of artificial intelligence in journalism. *Digital Journalism*, 10(10), 1756–1774.

Morse, J. M. (2012). The Implications of Interview Type and Structure in Mixed-Method Designs. In: Gubrium, J. F., Holstein, J. A., Marvasti, A. B., & McKinney, K. D. (Eds.). *The SAGE Handbook of Interview Research.* The Complexity of the Craft, 193–204.

Nagappan, N., Murphy, B., & Basili, V. (2008). The influence of organizational structure on software quality: An empirical case study. *Proceedings of the 13th International Conference on Software Engineering – ICSE '08*, 521.

Neff, G., & Stark, D. C. (2002). *Permanently Beta:* Responsive Organization in the Internet Era.

Neuberger, C., Nuernbergk, C. (2015). Verdatete Selbstbeschreibung der Gesellschaft: Über den Umgang des Journalismus mit Big Data und Algorithmen. In: Süssenguth, F. (Ed.). *Die Gesellschaft der Daten: Über die digitale Transformation der sozialen Ordnung*, 165–189.

Newman, N. (2018). *Journalism, media and technology trends and predictions 2018.* Reuters Institute for the Study of Journalism.

Newman, N., Fletcher, R., Robertson, C. T., Eddy, K., & Nielsen, R. K. (2022). *Digital News Report 2022*. Reuters Institute for the Study of Journalism & University of Oxford.

Nielsen, R. K., & Ganter, S. A. (2018). Dealing with digital intermediaries: A case study of the relations between publishers and platforms. *New Media & Society*, 20(4), 1600–1617.

Nguyen, A. (2013). *Online news audiences: The challenges of web metrics.*

Nguyen, D. B., Dow, C. R., & Hwang, S. F. (2018). An efficient traffic congestion monitoring system on internet of vehicles. *Wireless Communications and Mobile Computing*, 2018.

Nieborg, D. B., & Poell, T. (2018). The platformization of cultural production: Theorizing the contingent cultural commodity. *New Media & Society*, 20(11), 4275–4292.

O'Brien, H. L., Cairns, P., & Hall, M. (2018). A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies*, *112*, 28-39.

OpenAI (2022, November 30). Introducing ChatGPT. *OpenAI Blog*, https://openai.com/blog/chatgpt

Parasie, S., & Dagiral, E. (2013). Data-driven journalism and the public good: Computer-assisted-reporters and programmer-journalists in Chicago. *New Media & Society*, 15(6), 853–871.

Pavlik, J. V., & Bridges, F. (2013). The emergence of augmented reality (AR) as a storytelling medium in journalism. *Journalism & Communication Monographs*, 15(1), 4–59.

Pedersen, A.M., & Bossen, C. (2021). *Data Work in Healthcare:* An Ethnography of a BI Unit.

Pedersen, A. M. (2022). Making Reliable Data: Enacting and Negotiating Data Quality through Data Work. *Abstract from DASTS 2022*, Aarhus, Denmark.

Perreault, G., Kananovich, V., & Hackett, E. (2022). Guarding the firewall. *Journalism & Mass Communication Quarterly*.

Petre, C. (2015). *The traffic factories:* Metrics at chartbeat, gawker media, and the New York Times.

Petre, C. (2020). Engineering consent: How the design and marketing of newsroom analytics tools rationalize journalists' labor. In: *Measurable Journalism,* 121–139. Routledge.

Peuker, B. (2010). Akteur-Netzwerk-Theorie (ANT). In: Stegbauer, C., Häußling, R. (Eds.). *Handbuch Netzwerkforschung*, 325–335.

Picard, R. G. (2011). *The economics and financing of media companies* (2nd ed). Fordham University Press.

Platt, J. (2012). The History of the Interview. In: Gubrium, J. F., Holstein, J. A., Marvasti, A. B., & McKinney, K. D. (Eds.). *The SAGE Handbook of Interview Research*. The Complexity of the Craft, 9–26.

Plsek, P. E., & Greenhalgh, T. (2001). The challenge of complexity in health care. *Bmj*, 323(7313), 625–628.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of applied psychology*, 88(5), 879.

Polasky, S., Carpenter, S. R., Folke, C., & Keeler, B. (2011). Decision-making under great uncertainty: Environmental management in an era of global change. *Trends in Ecology & Evolution*, 26(8), 398–404.

Pollard, S. (1965). *The genesis of modern management: a study of the industrial revolution in Great Britain.*

Porlezza, C., & Splendore, S. (2019). From open journalism to closed data: Data journalism in Italy. *Digital journalism*, 7(9), 1230–1252.

Porter, T. M. (1996). *Trust in numbers: The pursuit of objectivity in science and public life.* Princeton University Press.

Postone, M. (1993). *Time, Labor and Social Domination:* A reinterpretation of marx's critical theory (1. ed). Cambridge University Press.

Power, D. J. (2007). *A brief history of decision support systems.*
Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2(3), 1–10.

Reventlow, N. J. (2020). Can the GDPR and Freedom of Expression coexist?. *AJIL Unbound*, 114, 31–34.

Robinson, J. A., & Acemoglu, D. (2012). *Why nations fail: The origins of power, prosperity and poverty.* Profile.

Rockwell, N. (2019). News in the Age of Algorithmic Recommendation. *Data Council*, 2019, Transcript.

Rosenthal, G., & Loch, U. (2002). Das Narrative Interview. In: Schaeffer, D., & G. Müller-Mundt (Eds.), *Qualitative Gesundheits- und Pflegeforschung*, 221–232.

Roszak, T. (1994). *The cult of information:* A neo-Luddite treatise on high-tech, artificial intelligence, and the true art of thinking. University of California Press.

Rothschild, A., DiSalvo, C., Johnson, B., Rydal B., DiSalvo, B. (2022). Interrogating Data Work as a Community of Practice. *Proceedings of the ACM on Human-Computer Interaction*, 6, Article 307 (November 2022), 29.

Ruckenstein, M., & Schüll, N. D. (2017). The datafication of health. *Annual review of anthropology*, 46, 261–278.

Ruppert, E., Law, J., & Savage, M. (2013). Reassembling social science methods: The challenge of digital devices. *Theory, Culture & Society*, 30(4), 22–46.

Sadowski, J. (2019). When data is capital: Datafication, accumulation, and extraction. *Big Data & Society*, 6(1), 2053951718820549.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3), 210–229.

Sanchez-Rola, I., Dell'Amico, M., Kotzias, P., Balzarotti, D., Bilge, L., Vervier, P. A., & Santos, I. (2019). Can I opt out yet? GDPR and the global illusion of cookie control. In: *Proceedings of the 2019 ACM Asia conference on computer and communications security*, 340–351.

Sarikaya, A., Correll, M., Bartram, L., Tory, M., & Fisher, D. (2018). What do we talk about when we talk about dashboards?. *IEEE transactions on visualization and computer graphics*, 25(1), 682-692.

Saunders, B., Kitzinger, J., & Kitzinger, C. (2015). Anonymising interview data: challenges and compromise in practice. *Qualitative Research*, 15(5), 616–632.

Schaetz, N. (2023). Journalism & Audience Datafication: How Audience Data Practices Shape Inequity. *Digital Journalism*, 1–21.

Schelling, T.C. (1978). *Micromotives and Macrobehavior.*

Schätz, K., & Kirchhoff, S. (2020). Neue Wege im Journalismus, Weichenstellung in der Ausbildung. *Journalistik*, Jg, 3, 98–110.

Schätz, K., & Pühringer, K. (2022). *Innovation durch Datafizierung:* Eine Erhebung des Einflusses von Digitalisierung und Datafizierung in österreichischen Tageszeitungen.

Schmandt-Besserat, D. (2014). The evolution of writing. *International encyclopedia of social and behavioral sciences*, 1–15.

Schmidt, C. (2019, July 25). How to cover 11,250 elections at once: Here's how The Washington Post's new computational journalism lab will tackle 2020. *Nieman Lab*, https://www.niemanlab.org/2019/07/how-to-cover-11250-elections-at-once-heres-how-the-washington-posts-new-computational-journalism-lab-will-tackle-2020/.

Schütze, F. (1983). Biographieforschung und narratives Interview. *Neue Praxis*, 13(3), 283-293.

Scott, W. R. (2004). Reflections on a half-century of organizational sociology. *Annu. Rev. Sociol.*, 30, 1–21.

Scott, M., Bunce, M., & Wright, K. (2019). Foundation funding and the boundaries of journalism. *Journalism Studies*, 20(14), 2034–2052.

Seawright, J., & Gerring, J. (2008). Case selection techniques in case study research: A menu of qualitative and quantitative options. *Political research quarterly*, 61(2), 294–308.

Selwyn, Neil (2022). Critical data futures. In: Housley, W., Fitzgerald, R., Beneito-Montagut, R., & Edwards, A. (Eds.). *Digital Society*, 593–609.
Shiryaev, A. N. (2016). *Probability-1* (Vol. 95). Springer.

Shoemaker, P. J., Vos, T. P., & Reese, S. D. (2009). Journalists as gatekeepers. In: Wahl-Jorgensen, K., Hanitzsch, T. (Eds.). *The handbook of journalism studies*, 93–107, Routledge.

Spee, A. P., & Jarzabkowski, P. (2009). Strategy tools as boundary objects. *Strategic organization*, 7(2), 223–232.

Splendore, S. (2016). Quantitatively oriented forms of journalism and their epistemology. *Sociology Compass*, 10(5), 343–352.
https://doi.org/10.1111/soc4.12366

Stake, R. E. (1995). *The Art of Case Study Research*. Sage Publications.

Stake, R. E. (2010). *Qualitative research: Studying how things work.* The Guilford Press.

Stalph, F. (2020). Evolving data teams: Tensions between organisational structure and professional subculture. *Big Data & Society*, 7(1), 205395172091996.

Star, S. L. (1989). The structure of ill-structured solutions: Boundary objects and heterogeneous distributed problem solving. *Distributed Artificial Intelligence*, 37–54.

Star, S. L., & Griesemer, J. R. (1989). Institutional Ecology, "Translations" and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, 19(3), 387–420.

Steenburgh, T. J., & Avery, J. (2010). Marketing analysis toolkit: Situation analysis. *HBS Case*, 510–079.

Stigler, S. M. (1986). *The history of statistics:* The measurement of uncertainty before 1900. Harvard University Press.

Strong, E. K. (1925). *The psychology of selling and advertising.* McGraw-Hill book Company, Incorporated.

Strong, E. W. (1951). Newton's Mathematical Way. *Journal of the History of Ideas*, 90–110.

Subramanian, A., & Nilakanta, S. (1996). Organizational innovativeness: Exploring the relationship between organizational determinants of innovation, types of innovations, and measures of organizational performance. *Omega*, 24(6), 631–647.

Subrahmanian, E., Monarch, I., Konda, S., Granger, H., Milliken, R., & Westerberg, A. (2003). Boundary objects and prototypes at the interfaces of engineering design. *Computer Supported Cooperative Work (CSCW)*, 12, 185–203.

Sullivan, J. L. (2019). *Media audiences:* Effects, users, institutions, and power. SAGE Publications, Incorporated.

Tandoc, E. C. (2014). Journalism is twerking? How web analytics is changing the process of gatekeeping. *New Media & Society*, 16(4), 559–575.

Tandoc, E. C. (2015). Why Web Analytics Click: Factors affecting the ways journalists use audience metrics. *Journalism Studies*, 16(6), 782–799.

Tandoc, E. C., & Thomas, R. J. (2015). The ethics of web analytics: Implications of using audience metrics in news construction. *Digital journalism*, 3(2), 243–258.

Tandoc, E. C. (2019). *Analyzing Analytics:* Disrupting Journalism One Click at a Time. Routledge.

Tandoc, E. C., & Thomas, R. J. (2015). The Ethics of Web Analytics: Implications of using audience metrics in news construction. *Digital Journalism*, 3(2), 243–258.

Taylor, L., & Broeders, D. (2015). In the name of development: Power, profit and the datafication of the global south. *Geoforum*, 64, 229–237.

Thurman, N., & Schifferes, S. (2012). The Future of Personalization at News Websites: Lessons from a longitudinal study. *Journalism Studies*, 13(5–6), 775–790. https://doi.org/10.1080/1461670X.2012.664341

Thurman, N., Moeller, J., Helberger, N., & Trilling, D. (2018). My Friends, Editors, Algorithms, and I. *Digital Journalism*, 7(4), 447–469. https://doi.org/10.1080/21670811.2018.1493936

Tong, J. (2022). *Data for journalism:* Between transparency and accountability. Taylor & Francis.

Tong, J., & Zuo, L. (2021). The inapplicability of objectivity: Understanding the work of data journalism. *Journalism Practice*, 15(2), 153–169.

Tourangeau, R., Rips, L. J., Rasinski, K. (2000). *The Psychology of Survey Response.* Cambridge University Press.

Trinczek, R. (2009). How to Interview Managers? Methodical and Methodological Aspects of Expert Interviews as a Qualitative Method in Empirical Social Research. In: Bogner, A., Littig, B., & Menz, W. (Eds.). *Interviewing Experts.* 203–216.

Trudeau, R. J. (2013). *Introduction to graph theory*. Courier Corporation.

Turow, J. (2017). *The aisles have eyes: How retailers track your shopping, strip your privacy, and define your power.* Yale University Press.

Van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society*, 12(2), 197–208.

Van Drunen, M. Z., Helberger, N., & Bastian, M. (2019). Know your algorithm: what media organizations need to explain to their users about news personalization. *International Data Privacy Law*, 9(4), 220–235.

Van der Haak, B., Parks, M., & Castells, M. (2012). The future of journalism: Networked journalism. *International Journal of Communication*, 6(16), 2923–2938.

Van Schalkwyk, F., Willmers, M., & McNaughton, M. (2016). Viscous open data: The roles of intermediaries in an open data ecosystem. *Information Technology for Development*, 22(sup1), 68–83.

Vulpius, J. (2022). The Role of Audience Data and Metrics in Becoming 'Audience-Centred'. In: V. J. E. Manninen, M. K. Niemi, & A. Ridge-Newman (Eds.), *Futures of Journalism*, 347–363. Springer International Publishing.

Waisbord, S. (2019). *Communication:* A post-discipline. John Wiley & Sons.

Wallace, J. (2018). Modelling contemporary gatekeeping: The rise of individuals, algorithms and platforms in digital news dissemination. *Digital Journalism*, 6(3), 274–293.

Wassermann, S. (2015). Das qualitative Experteninterview. In: Niederberger, M., Wassermann, S. (Eds.) Methoden der Experten- und Stakeholdereinbindung in der sozialwissenschaftlichen Forschung. Springer VS.

Weber, J., & Schäffer, U. (2008). *Einführung in das Controlling*. Schäffer-Poeschel.

Weikart, R. (2016). *From Darwin to Hitler: evolutionary ethics, eugenics and racism in Germany*. Springer.

Westlund, O., & Lewis, S. C. (2014). Agents of media innovations: Actors, actants, and audiences. *The Journal of Media Innovations*, 1(2), 10–35.

Williams, A., Miceli, M. & Gebru, T. (2022, October 13). The Exploited Labor Behind Artificial Intelligence. Supporting transnational worker organizing should be at the center of the fight for "ethical AI". *Noema Magazine*. https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/.

Willig, I. (2022). From audiences to data points: The role of media agencies in the platformization of the news media industry. *Media, Culture & Society*, 44(1), 56–71.

Wilson, D. C. (2014). Arnold Toynbee and the industrial revolution: The science of history, political economy and the machine past. *History & Memory*, 26(2), 133–161.

Wing, J. M. (2019). The Data Life Cycle. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.e26845b4

Wu, S., Tandoc Jr, E. C., & Salmon, C. T. (2019). When journalism and automation intersect: Assessing the influence of the technological field on contemporary newsrooms. *Journalism Practice*, 13(10), 1238–1254.

Wunderman, L. (1994). Interactive communications, the dream and the reality. *Journal of Direct Marketing*, 8(3), 74–77.

Yankelovich, D. (1971). Interpreting the New Life Styles. *Sales Management, the Marketing Magazine*, Nov. 16, 1971, 26–28.

Yin, R. K. (2018). *Case study research and applications: design and methods, 6e*. Thousand Oaks, Sage.

Zak, P. (2013, July 4). Measurement Myopia, *The Drucker Institute Archive*, https://drucker.institute/thedx/measurement-myopia/.

Zahid, H., Mahmood, T., Morshed, A., & Sellis, T. (2019). Big data analytics in telecommunications: literature review and architecture recommendations. *IEEE/CAA Journal of Automatica Sinica*, 7(1), 18–38.

Zamith, R. (2018). Quantified audiences in news production: A synthesis and research agenda. *Digital Journalism*, 6(4), 418–435.

Zhan, Y., Wan, P., Jiang, C., Pan, X., Chen, X. & Guo, S. (2020). Challenges and Solutions for the Satellite Tracking, Telemetry, and Command System, *IEEE Wireless Communications*, 27(6), 12–18.

Zuiderveen Borgesius, F., Trilling, D., Möller, J., Bodó, B., De Vreese, C. H., & Helberger, N. (2016). Should we worry about filter bubbles?. *Internet Policy Review. Journal on Internet Regulation*, 5(1).

8.4 Abstract

In recent years, the role of data at digital news publishers has expanded beyond the applications of editorial analytics and data journalism—yet research on data in journalism mostly centers around editorial concerns. Starting from this observation, I look outside the newsroom to explore the overall impact of data and data work in the field. With this thesis, I aim to understand how metrics, data and data-driven thinking contribute to organizational and structural transformation and how they might redefine professional boundaries within news organizations. Based on a comprehensive theoretical framework that encompasses data work as an analytical category, I employ a qualitative research design with expert interviews from six case studies conducted at publishers in Germany.

Among my findings, I conclude how fundamental changes in data practices, organizational structure, and management culture at these organizations took place in recent years. New roles have emerged, creating a novel professional class of data workers organized into newly established data departments that operate independently from editorial. This shift in data practices can be attributed to a re-orientation towards subscriptions and individual customers. As the organizations experiment with various data approaches and metrics, neither the metrics nor the associated strategies have stabilized, yet the practices among individual data workers remain relatively consistent across the sample. A normative shift could be considered complete in the recreation of analytics infrastructure, dashboarding, and reporting affordances, and the "making up" of metrics by the publishers, two functions which were formerly handled by external analytics providers. I also show how artifacts such as dashboards, reports and metrics are not neutral either as they marginalize editorial power through the renegotiation of boundaries.

On the question of organizational isomorphism, a certain degree of uniformity of outcomes can be observed across the sample. There are indicators of normative pressure in the shape of data bureaucratization, with data workers at the news organizations establishing a cognitive basis for their occupational autonomy through metrics and measurement infrastructure, while mimetic and coercive isomorphisms are found to a lesser extent. I also demonstrate that experimentation with data and data infrastructure takes place regardless of company size. Furthermore, the particular data imaginary adopted by an organization bears no clear correlation (either positive or inverse) with editorial power.

Overall, I point out the emergence of a new class of data professionals who are certainly influential within their organizations and are able to shape key metrics, provide analytical insights, and justify managerial decisions. While these developments underscore a broader transformation within the field, the effects of their work remain somewhat unclear. With this thesis, I hope to not only identify current trends and changes in journalism but also provide input for future research into the pervasive role of data work in other industries.

German translation follows

In den letzten Jahren hat sich der Stellenwert von Daten bei digitalen Nachrichtenverlagen über Analytics und Datenjournalismus hinaus stark entwickelt—dennoch konzentriert sich die Forschung über Daten im Journalismus hauptsächlich auf redaktionelle Belange. Ausgehend von dieser Feststellung blicke ich in die Unternehmen hinter den Redaktionen, um die allgemeinen Auswirkungen von Daten und Datenarbeit im Feld zu untersuchen. Mit dieser Arbeit möchte ich verstehen, wie Metriken, Daten und datengetriebenes Denken zu organisatorischen und strukturellen Veränderungen beitragen und sich professionelle Grenzen innerhalb von Nachrichtenorganisationen verschieben. Auf der Grundlage eines theoretischen Bezugsrahmens, der Datenarbeit als analytische Kategorie umfasst, verwende ich ein qualitatives Forschungsdesign aus Experteninterviews und Fallstudien in deutschen Verlagen.

Unter anderem folgere ich, dass sich Daten- und Managementpraktiken in den untersuchten Organisationen in den letzten Jahren grundlegend verändert haben. Es haben sich neue Rollen herausgebildet, die eine neue professionelle Klasse von Datenarbeitern bilden, oftmals in neu eingerichteten Datenabteilungen organisiert und relativ unabhängig von den jeweiligen Redaktionen. Dieser Wandel in der Datenpraxis lässt sich auf eine Rückbesinnung auf Abonnements und Einzelkunden zurückführen. Während die Unternehmen mit verschiedenen Ansätzen und Metriken experimentieren und sich dahingehend noch nicht stabilisiert haben, bleiben die Praktiken der einzelnen Datenarbeiter in der Stichprobe relativ einheitlich. Eine normative Verschiebung kann insofern als vollzogen gelten, als dass sich der Betrieb von Datenstrukturen und die Definition von Metriken nunmehr in den Verlagen abspielen—Funktionen, die zuvor von externen Dienstleistern übernommen wurden.

Zusätzlich kann gezeigt werden, dass Artefakte wie Dashboards, Berichte und Metriken nicht als neutral zu begreifen sind, da sie die redaktionelle Macht durch eine Neuverhandlung von Grenzen marginalisieren. Was die Frage der organisatorischen Isomorphie betrifft, so ist in der Stichprobe eine gewisse Einheitlichkeit der Ergebnisse zu beobachten. Es gibt Indikatoren für normativen Druck in Form von Datenbürokratisierung, wobei die Datenarbeiter in den Nachrichtenorganisationen eine kognitive Grundlage für ihre berufliche Autonomie durch Metriken und Messinfrastrukturen schaffen, während mimetische und erzwingende Isomorphismen in geringerem Ausmaß zu finden sind. Ich zeige auch, dass das Experimentieren mit Daten und entsprechenden Technologien unabhängig von der Unternehmensgröße stattfindet. Darüber hinaus scheint keine eindeutige Korrelation (weder positiv noch umgekehrt) zwischen den jeweils angenommenen Datenimaginationen und dem Grad der redaktionellen Einflussnahme zu bestehen.

Insgesamt beschreibe ich eine neue Klasse von Datenexperten, die in ihren Unternehmen durchaus einflussreich darin sind, wichtige Kennzahlen zu gestalten, analytische Erkenntnisse vorzubereiten und Management-entscheidungen zu rechtfertigen. Während diese Entwicklungen einen breiteren Wandel innerhalb des Journalismus unterstreichen, bleiben die Auswirkungen der Datenarbeit schwer zu greifen. Neben einiger Desiderata möchte ich mit dieser Arbeit nicht nur aktuelle Trends und Veränderungen im Journalismus aufzeigen, sondern auch Anregungen für künftige Forschung über die allgegenwärtige Rolle der Datenarbeit in anderen Branchen geben.