Developing Data Stories in Digital Humanities: challenges and protocol

Willemien Sanders[1], Roeland Ordelman[2], Mari Wigham[2], Rana Klein[2], Jasmijn van Gorp[1], and Julia Noordegraaf[3]

[1]Utrecht University
[2]Netherlands Institute for Sound and Vision
[3]University of Amsterdam

This article discusses the development of data-driven stories and the editorial processes underlying their production. Such 'data stories' have proliferated in journalism but are also increasingly developed within academia. Within CLARIAH, the Common Lab Infrastructure for the Arts and Humanities, we are developing data stories based on analyses of data and metadata available via the Media Suite, an online resource providing access to a wide range of multimedia collections. Although 'data stories' lack a clear definition, there are similarities between the processes that underlie journalistic and academic data stories. However, there are also differences, specifically when it comes to epistemological claims. In this article we discuss data stories as phenomenon and their use in journalism and in the Humanities, based on the three main elements of data stories: data, visualisation, and narration. This provides the context in which we developed an editorial protocol for the development of CLARIAH Media Suite Data Stories, which includes four phases: exploration, research, review, and publication. While exploration focuses on data selection, research focuses on narration. Visualisation plays a role in both of these phases. Review is geared towards quality control, and in the publication phase the data story is published and monitored. By discussing our editorial protocol, we hope to contribute to the debate about how to develop and account for academic data stories.

1 Introduction

The datafication of society and – more specifically – of cultural heritage collections has continued to thrive over the past two decades and has facilitated the development of a flourishing Digital Humanities. Collection owners have developed infrastructures to make digitised or born digital collections accessible, and to allow activities such as the searching, bookmarking, annotating and analysing of collection content. One example of such a Digital Humanities infrastructure is the CLARIAH Media Suite.[1]

[1] See `https://mediasuite.clariah.nl`.

Within CLARIAH, the Common Lab Infrastructure for the Arts and Humanities, the Media Suite has been developed for and with media and humanities scholars to facilitate searching and working with the collections it holds (Melgar et al., 2017; Melgar-Estrada et al., 2019; Ordelman et al., 2018; Van Gorp et al., 2021). The Media Suite provides access to multimedia collections, including, among others, large bodies of historical radio, television, film, and newspaper data. It currently offers 92 data sets from 62 cultural heritage and knowledge organisations in the form of both data and metadata. It is expanded continuously as Dutch public broadcasting material in added on a daily basis. The Media Suite is hosted by the Netherlands Institute for Sound and Vision (NISV). Researchers can access the Media Suite via their institutional login, but the majority of the data is not publicly available due to copyright restrictions.

The media and culture studies scholars involved in the development of the Media Suite are trained primarily in qualitative methods (Van Gorp et al., 2021). However, the large data sets that the Media Suite provides access to are also a valuable potential source for quantitative research. As Windhager et al. (2019) argue, the growing digital collections invite new ways of analysing and understanding those collections. While qualitative research offers opportunities for close reading and the interpretation of individual items, quantitative research allows the discovery of patterns and possible changes, e.g., over time, by distant reading large bodies of work.

The term 'data stories' has been used in a variety of publications to refer to different narratives involving quantitative data and, usually, visualisations, especially in journalism (Stalph and Heravi, 2021; Weber, 2020). According to Kennedy and Engebretsen (2020), data-based technologies have come to play a growing role in informing and persuading the audience. El Outa et al. (2022) observe that there is a lack of integrated tools for making data stories. In the Digital Humanities there are a number of initiatives that aim to publish (data-based) research online in a format that addresses the demands of Humanities scholarship, sometimes referred to as enhanced publications: publications enriched with or linked to related research results (Bardi and Manghi, 2015). For instance, Scalar is a platform for enriched multimedia publications, facilitating multi-layered publications that might include both the central argument as well as links to the underlying material it is based on, through the import of metadata.[2] The Journal of Digital History is characterised by "multilayered publication" for data-driven research and "transmedia storytelling", as different layers with different media can be included.[3] Related initiatives include the Journal of Open Humanities Data, which focuses on the sharing and reuse of data and techniques, and the Journal of Data Mining & Digital Humanities, which positions itself at the intersection of computing and the Humanities.[4]

Given that much of the content accessible via the Media Suite is copyright protected and using other platforms for research and publication is therefore not an option, we have started experimenting with data-driven research and publications ourselves: the Media Suite Data Stories.[5] These data stories can be understood as an additional layer on top of the Media Suite infrastructure, an integrated layer which facilitates analyses and the publication of – mainly – quantitative research with the Media Suite data and metadata. In addition, and making this a rather unique form of publication, the underlying aggregated data are accessible via interactive visualisations, providing an

[2] See `https://scalar.me/anvc`.

[3] See `journalofdigitalhistory.org`.

[4] See `openhumanitiesdata.metajnl.com` and `https://jdmdh.episciences.org`.

[5] See `https://mediasuitedatastories.clariah.nl` (predominantly in Dutch).

opportunity for readers to scrutinise the authors' works.

El Outa et al. (2022) observe that there is a lack of complete descriptions of workflows for making data stories. Figueiras and Vizoso (2022) also call for guidelines for the iterative process of making 'visual data stories'. Fickers and Clavert (2021, par.10), reflecting on their experiences in developing the Journal of Digital History, argue that "we are in desperate need for clearer protocols on how to document and share self-reflexive position statements in scientific publications online". Although referring to historians, the same can be said about scholars in other domains: we all need to develop ways of documenting and discussing our interactions with digital and online research and publication technologies. We hope to contribute to this discussion with this article.

According to Riche et al. (2018, p.5), the development of data-driven stories occurred in both academia and journalism, although the former tends to interrogate the possibilities and potential of visualisations while the latter are "at the forefront" of the production of data-driven stories. So far, Media Suite Data Stories have been produced by both scholars and journalists. Therefore, in this article, we focus on scholars and journalists to discuss data stories. We argue that there are specific demands for the production process of a data story, which we have documented in an editorial protocol for making Media Suite Data Stories.

In what follows we will discuss the main elements of data stories, data, visualisation and narration, in the context of journalism and scholarship to create the context for our discussion of the editorial protocol for Media Suite Data Stories. This editorial protocol aims to safeguard the quality of both the content of and technology used for creating Media Suite Data Stories. We conclude with ideas for future work.

2 Data

There is a popular idea that "some of today's most relevant stories are buried in data" (Riche et al., 2018, p. 4; see also Figueiras and Vizoso, 2022). Considering the problem with data bias and the challenges of understanding data properly (of which power imbalance between those usually collecting data and those usually represented in data is not the least one (D'Ignazio and Klein, 2020)), ideas like these need critical scrutiny.

According to Arrese (2022), the use of data for journalism can be traced back to early economic journalism in the 17th and 18th centuries, when prices of goods and stocks started to be regularly reported. This developed into data-driven reporting in the 19th century, for instance, of financial data. With the growing belief in numbers and measurements as well as the development of digital infrastructures and services, data have become increasingly popular in other fields as well, and so journalists widened their scope to report data-based news in a wide variety of fields.

Although data are often regarded as objective reflections of certain situation, as "descriptions of a priori conditions" (Drucker, 2011, par.1) or even as the basis for 'facts' (see, for instance, Cairo, 2019; El Outa et al., 2020), Drucker (2011) urges us to think about data as interpretations, given their constructed nature. She calls for an understanding of data as capta, 'taken', since statistics are a form of rhetoric: they are a way of talking about a reality, not reality itself. Schöch (2013, n.p.) likewise defines data in the humanities as "a digital, selectively constructed machine-actionable abstraction representing some aspects of a given object of humanistic inquiry". We recognise the importance of the distinction between 'capta' and data. However, we stick to the terms 'data' and 'metadata' as a proper discussion of this critical perspective

is beyond the scope of this article and we fear using capta might be confusing rather than clarifying.

With respect to journalism, Weber et al. (2018, p.191) discuss two definitions of data journalism. The first relies on work by Howard (2014): "gathering, cleaning, organizing, analyzing, visualizing and publishing data to support the creation or acts of journalism". The second relies on work by Rinsdorf and Boers (2016): "analyzing open data sets using (semi-)automatized methods to detect meaningful patterns in data structure". Gray and Bounegru (2021, p.3) limit their explanation to "collecting, analysing, visualizing and narrating data". Although 'cleaning and organising' might be understood as forms of exploring data, the general idea that these authors seem to voice is that the data sets they are referring to are more or less ready for use.

D'Ignazio and Klein (2020) point to the need for contextualisation of data: information about how and why data are collected and how they are processed - in other words, how data are "produced" (Kitchin, 2021, p.5) - before they are turned into a data set ready for use and into visualisations (Kennedy and Engebretsen, 2020).

Journalists partly rely on their own data collection, for instance, of social media data, but according to Stalph and Heravi (2021) as well as El Outa et al. (2022), they greatly rely on open data and data provided by organisations, including government agencies. Those data, before being offered to outsiders to work with, have been collected and processed. This includes, amongst others, parsing, filtering, and refining (Fry, 2008). Outliers and missing data might also have been dealt with, amongst others.

Such public data is often offered without context, or with only minimal context, as governments support the sharing of data but allocate insufficient resources to provide proper context about the collection and processing (or cleaning up) of the data (D'Ignazio and Klein, 2020). Therefore, it is hard to properly understand such data and interpret the results of analyses. Such data acquire a "quiet authority" (Lowrey et al., 2019, p.70). Ettema and Glasser (2006) argue that the knowledge claims that are based on such pre-processed data sets are pre-justified by the organisation that provided those data, taking the weight of accounting for them off the shoulders of journalists. In addition, Stalph and Heravi (2021, p.20) argue that an "overreliance on data ... from governmental sources" might lead to the perpetuation of hegemonic discourses. In such cases, journalists reinforce rather than monitor the powers that be.

For Humanities scholars it is vital that they are clear about their research practices and activities. On the one hand, journalists use data research mainly to uncover more or less hidden truths for a general audience. These truths are believed to be unaffected by journalists' values (Parasie, 2015), and as discussed above, are often based on publicly available data sets, and thereby the readers approach the results of the analysis of data as reflecting a truth that speaks for itself. On the other hand, Humanities scholars are used to constructing an argument, a convincing narrative from a position of situated knowledge (Drucker, 2011; Haraway, 1988) and the results of data analyses are extensively contextualised to make the argument convincing and reliable and for it to be reproducible by peers.

Fickers and Clavert (2021, par.16) discuss the "new digital practice of writing, visualizing, and arguing history" in their editorial to the first issue of the Journal of Digital History. This practice might include "producing transparency about how the digital interferes in the iterative process or lifecycle of the research process" which may be considered an "epistemological imperative". Transparency about the role of digital technology in the process of data production and analysis is demanded for methodological clarity. Bardi and Manghi (2015) argue that, as publications are only

released at the very end of the research process, data and their processing methods should be shared as well in order to meet academic standards of replicability. In other words: the digital scholarly practice should include transparency with respect to how scholars include the digital data and tools they use in their research practices (Condie and Costa, 2018), because such new digital practices have consequences for the epistemological claims scholars may present and that result from their situated position.

Media Suite data are complex and messy. As the Media Suite provides access to a wide range of different (historical) resources, the data are extremely heterogeneous. Also, data fractures occur and data might be missing. Fractures and missing data occur as a result of changing collection and metadata policies and practices. For instance, originally archivists at NISV added metadata to the television and radio archive. Since a number of years, media professionals at public broadcaster NPO do this. In addition, data might be missing or be unfindable due to technical hick-ups, collection policies, and human mistakes. Examples include errors in Optical Character Recognition (OCR) and Automatic Speech recognition (ASR) transcripts, typos and the limited and purposeful use of expensive and/or time consuming technologies, such as face and voice recognition tools.

The data sets available in the Media Suite come with extensive explanation, offered through the Media Suite data registry.[6] For each collection the data registry provides information about its size, the kinds of media it includes and other characteristics. This gives users a minimum of necessary information about the data available and thus contextualises the data somewhat, as advocated by D'Ignazio and Klein (2020). Unfortunately, as data processing from, for instance, the national public broadcaster NPO to the Media Suite remains a black box for many, creating full transparency about the data is not a matter of course. For example, the reasoning of NPO staff when adding specific metadata is not always evident, but highly relevant as interpretative act (Drucker, 2011). As a consequence, when creating a corpus, it is necessary to explore the desired data to understand what is there and what is missing. That is why exploration is the first step in our editorial protocol.

According to Condie and Costa (2018, p.206), it is crucial to address how projects "recognize the interplay between technology and research". From the very beginning, visualisations play a role in the editorial process of making a Media Suite Data Story. Below we discuss visualisation further.

3 Visualisation

Visualisation is rooted in cartography, developed centuries ago, and continued to be used in planning and commerce (Kennedy and Engebretsen, 2020). Visualisations nowadays play an important role in communicating data-based insights. Stalph and Heravi (2021) argue that data visualisations have both an epistemological and a communicative function. Journalists, like scholars, first explore their data sets and later visualise their findings to both inform journalistic claims and to communicate these claims (Stalph and Heravi, 2021). Weber et al. (2018, p.192) argue that data visualisations are central to data stories, visualisations "that range from simple to complex multi-modal interactive stand alone graphics". Visualisations come in many shapes and forms, which all carry meaning. Cairo (2019) observes they are seductive, but they are often also used selectively to meet the interests of authors.

[6] See `https://mediasuitedata.clariah.nl`.

According to Riche et al. (2018, p.8), data-driven stories "start from a narrative that either is based on or contains data and incorporates this data evidence ... to confirm or augment a given story", often in the form of visualisations. Data visualisations help to take the reader through the information and knowledge gained from the data.

Following McCosker and Wilken (2014), Stalph and Heravi (2021) refer to the process of creating data visualisations as 'diagrammatic thinking'. Considering data visualisations as a combination of elements, diagrammatic thinking seeks to shape and limit the potentially unlimited number of possible combinations of elements and thereby what stories can be told. From this perspective, visualisations can be understood as presenting a problem rather than an answer: they invite discussion of what they represent rather than serve as answers.

Drucker's 2011 approach to data discussed above has repercussions for our understanding of data visualisations as well. Drucker argues humanistic knowledge relies on interpretation. As a consequence, visualisations express such interpretations. At the same time, visualisations hide that interpretation behind 'transparent' graphs. A shift to graphs that show such interpretations is essential for Drucker, but so far this approach seems unpopular among practitioners, as Stalph and Heravi (2021) found that conventional static bar charts are still used widely. Also, there is a tendency to focus on visualizations in terms of efficiency and 'correctness' (see for instance Cairo, 2019; Kennedy and Engebretsen, 2020), rather than on their rhetoric, aesthetics and/or affect (see D'Ignazio and Klein, 2020).

Discussing data visualisations in journalism, Weber et al. (2018) argue that data visualisation calls for transparency regarding practices and editorial processes, both as qualitative management strategy and as ethical standard. This includes explaining the collection, analysis, and presentation of data and allowing users to check these practices.

Duangphummet and Ruchikachorn (2021) share their lessons learned in different phases of developing various data visualisations. While working on a series of visualisations they updated the interdisciplinary team. It was only after the first iteration that they included a domain expert in the team. Later they also involved a data scientist and analyst.

Although so far we have not created complex visualisations consisting of many elements, we treat the interpretation of visualisations in Media Suite Data Stories never as a matter of course. Rather, they are contextualised with respect to the available data and their overall context, and they are usually a starting point for further research, analysis, and discussion. In that sense they are also a "procedural tool" (Stalph and Heravi, 2021, p.5) in the investigation itself: during the exploration phase they mainly help to understand the data and during the research phase they mainly communicate findings. In addition, to ensure the interpretation and data story make sense, we involve a domain expert in the development of each Media Suite Data Story.

With respect to their form, and based on the work of a number of other scholars, Weber et al. (2018, p.192) describe data stories as multi-modal and hybrid artefacts that can be based on numbers, texts and images to create "a coherent whole". Both definitions of data journalism by Weber et al. (2018) discussed above focus on data and their analysis, while Gray and Bounegru (2021) also refer to narrating (see Data). For Media Suite Data Stories visualisations play a role in both exploration and research narration. That is why we discuss narration next.

4 Narration

Most work on data visualisations in journalism discusses visualisations as a way to tell stories. By including a narrator, sequentiality, time and "tellability" (Weber, 2020, p.307), data visualisations become narratives. To understand how journalists make use of data, various scholars have analysed journalistic 'data stories'. For instance, Segel and Heer (2010) have analysed 58 data-based stories and visualisations and provided a taxonomy including seven genres: magazine style, annotated chart, partitioned poster, flow chart, comic strip, slide show, and film/video/animation. Each can be positioned on an axis between author-driven and reader-driven narratives. The former are more explanatory in nature, the latter more exploratory (see also (Weber, 2020).

Further developing this work, Stolper et al. (2016) distinguish between four broad non-exclusive categories: communicating narrative and explaining data; linking separated story elements; enhancing structure and navigation; and providing controlled exploration. Evidently, this includes a wide variety of ways in which data can be visualised and stories can be told. Stalph and Heravi (2021) more recently analysed 185 visualisations with a focus on testing a synthesised framework for analysis, which includes visualisation types, interactivity, data sources, data access and purpose. Of the 156 that could be classified based on their visualisations, "the majority" (Stalph and Heravi, 2021, p.15) used some narrative device, such as a temporal display, showing changes over time.

Based on the above we define 'data stories' as follows: Data stories are the output of an iterative process of data collection, data exploration, data preparation, data analysis, data visualisation, interpretation and narration. The result is a narrative that is based on the analysis of quantitative data and includes visualisations of these data with explanations of their production, meaning and their context. In addition, we might distinguish between the practice of storytelling *through* visualisations, which seems more applicable to current practices in journalism, and storytelling *with* visualisation, which seems more applicable to current practices in the Humanities and reflects both work published in the Journal of Digital History and Media Suite Data Stories. In the latter, the visualisations are part of a larger story in writing that explains and interprets the research the story is reporting. Narration happens by explaining and interpreting the findings, and is therefore closely tied-up with our research phase.

The above sketches the context within which we are developing the Media Suite Data Stories as a publication infrastructure on top of the Media Suite. Below we discuss our editorial workflow within this context.

5 The Media Suite Data Stories editorial protocol

Media Suite Data Stories combine visualisations of data analyses with texts that explain and interpret those visualisations and connect them into a narrative. Fickers and Clavert (2021) argue the potential of opening up the research process to readers by providing not only the results of a given research project but also the process itself. The "weaving [of] interpretation, narrative, evidence, and commentary" (Fickers and Clavert, 2021, par.4) allows the reader to "think along" with the goal to be transparent about the way digital technologies interact with the research process, as this has epistemological consequences. Thinking along is what we aim to facilitate with our Media Suite Data Stories.

Media Suite Data Stories are characterised by (a) the combination of domain-specific

knowledge and a critical approach to data and tools, and (b) transparency with respect to data exploration and analysis. Although the data in the Media Suite are largely copyright protected, the results of analyses or aggregated data are accessible to readers. At the same time, in the Media Suite Data Stories, the production (e.g., through ASR, voice recognition, and face recognition), exploration, selection, processing, analysis and visualisation of data are all accounted for in methodological accounts and available for scrutiny. Media Suite Data Stories are produced with the help of Jupyter Notebooks. Unfortunately, again due to copyright restrictions on the data in the Media Suite, these are not yet widely available.

Media Suite Data Stories also align with what Stolper et al. (2016, p.1) refer to as "author-defined" narratives. They are published as web pages and the reader can scroll down as they progress through the story. Publishing stories this way is also referred to as 'scrollytelling' (Weber, 2020). Stolper et al. (2016, p.12) argue that scrolling is "a pervasive and powerful technique used in data-driven storytelling" even if a better understanding of its strengths and weaknesses is needed.

The first Media Suite Data Story we published functioned as a 'proof of concept' and a showcase of a variety of techniques to visualize characteristics of the then-popular TV show *De Wereld Draait Door*, rather than a profound story about it.[7] It proved promising enough to set up a team to further develop Media Suite Data Stories.

In order to understand what it means to develop a data story, we started by developing another Media Suite Data Story ourselves. We included three data science students in the team to help us analyse the presence of female and male nouns and pronouns, such as she, her, miss, and madam and their male equivalents, to understand the presence of women and men in Dutch television and radio. The students presented their work at their university and a political scientist was interested in using this technology to analyse the media presence of politicians during the then-upcoming parliamentary elections. This eventually led to a data story on the presence of female and male politicians in the media during the six weeks preceding the 2021 parliamentary elections in the Netherlands.[8].

Based on these experiences, we wrote a first version of our editorial workflow, which we later refined based on additional experiences, and which includes four (iterative) stages. It is aimed at safeguarding the quality of both the domain-specific knowledge transfer and a critical approach to digital data and tools, as well as transparency of the research process and outcomes.

According to Fickers and Clavert (2021, par.10), workflows describe the experimental process of knowledge production in a formalized way. By providing insight into the editorial protocol, we not only hope to develop the discussion by Fickers and Clavert (2021) as well as Duangphummet and Ruchikachorn (2021) concerning digital research and publication practices referred to above, but also take a step towards lowering the threshold for prospective researchers interested in conducting and publishing research based on multi-modal media data and metadata.

Below we discuss each of the four phases in the editorial protocol: exploration, research, review, and publication.

---

[7] See `https://mediasuitedatastories.clariah.nl/DWDD`.

[8] See `https://mediasuitedatastories.clariah.nl/elections-2021-first-results` and `https://mediasuitedatastories.clariah.nl/elections-dec-2021`.

5.1 Exploration

Media Suite Data Stories are based on - and their production is driven by - a research question or hypothesis. This is because, as discussed above, the available data in the Media Suite is too extensive and diverse to just simply rely on 'story finding'. In addition, an open exploration to find a story runs the risk of finding different results and connecting them without justification. Therefore, we use exploration to develop a research question or hypothesis and test a preliminary data set, to understand those data and their potential for answering the question or hypothesis.

To tell stories based on data, it is necessary to be or become acquainted with those data and understand what they represent. An exploration of the data that goes beyond 'cleaning' is part of any research effort. Such exploration should be geared towards understanding what the data are about, what the possibilities and limitations are, and which analyses seem fruitful. According to Schöch (2013) this need is a characteristic of 'big data' (however "fashionable" the concept (Kennedy and Engebretsen, 2020, p.22)). An AI specialist and a data engineer (to whom we refer as data scientists) are part of the editorial team of Media Suite Data Stories to guarantee that knowledge about the (meta)data is present in each project.

Four steps are taken in this phase. First, the researcher and data scientists brainstorm about the research interest and potentially useful data. For example, for our data story on the 2021 parliamentary elections, we selected programmes in which guests appear, which includes news and current affairs but also talk shows and interview programmes. We assumed it was likely politicians would show up in such programmes during the campaign weeks. We excluded programmes on, amongst others, culture and travel, and music shows as we expected they would be rather irrelevant. In the MediaOorlog (MediaWar) project the researchers prepared their data set by manually enriching the Media Suite with metadata with different categories of newspapers published during the war, distinguishing between those published or controlled by the nazis and anti-nazi newspapers (illegal press and those published in liberated areas). This provided the opportunity to compare these publications in ways that were not possible before.[9] In this case, data preparation included the further categorisation of the selected data.

Second, the researcher and data scientist translate their ideas into measurable questions, suitable for quantitative research. For the data story on the elections, the guiding questions focused on which politicians were present in the media and, by extension, which parties received media attention. Using face and voice recognition on our data set, we timed their appearances to measure this.

For the Media Suite Data Story on the 'fake news' discourse on Dutch television the author searched for a number of related terms, including fake news, nepnieuws (the Dutch translation of fake news), fake berichten or nepberichten (fake messages), misinformatie (misinformation), desinformatie (disinformation) en nepinformatie (fake information) to represent the discourse.[10]. He then counted the number of programmes in which these terms appeared in the subtitles.

Third, the researcher and data scientists execute a preliminary research project, like a pilot study, to test their ideas and data set. And fourth, they discuss the results to see if they can proceed to the next phase. Our project on the investigative programme *Tegenlicht* (literally: Backlight) provides a good example here. In 2021-2022 we developed this data story to mark the 20th anniversary of the programme. We used the transcripts

[9] See `https://mediasuitedatastories.clariah.nl/mediaoorlog`.

[10] See `https://mediasuitedatastories.clariah.nl/fake-news-2023`.

of the episodes as a data set (transcripts of a few episodes from the early days were missing though). With the *Tegenlicht* senior researcher and one of the editors in chief we initially wanted to focus on 'climate', given its topicality. However, initial analyses of the presence of the term climate in the transcripts yielded no noteworthy results: visualisations showed no remarkable developments over the years in the occurrence of the term. Also, word clouds based on the transcripts of episodes that dealt with climate did not result in anything worth pursuing. An explanation might be that the term 'climate' is also used in combination with, for instance, the economy or culture. We decided to drop the exploration and focus on another idea: technology. [11] This eventually resulted in the published Media Suite Data Story.

All phases in the Media Suite Data Story process are iterative, and the team might need several iterations to adjust and improve the question/hypothesis and data set to come to a coherent and feasible research plan and accompanying data. This process helps to get from 'big data' to 'smart data', which are limited in size, (semi) structured and rather heterogeneous and can be analysed using specific tools and methods (Schöch, 2013). In other words: smart data are more focused towards a specific task.

Once the team feels the research question/hypothesis and data set match sufficiently and provide promising first results, they move to the second phase: the research phase.

5.2 Research

The research phase is aimed at properly investigating the research question/hypothesis and narrating a Media Suite Data Story. In this phase, the focus is on data analysis and interpretation through visualisations. The complete data set is selected based on the exploration phase. The progress of the research depends on the results of consecutive steps. That is, based on the first analyses and visualisations thereof, patterns or developments may occur that deserve further investigation. As a result, in this phase, the narrative of the story is also constructed. To get there, the team takes five steps in this phase.

First, they create focus in their project by specifying the questions/hypotheses to further investigate the data set. Second, they further operationalise these questions/hypotheses so that they can be investigated. Obviously, some overlap with the Exploration phase occurs here, but the Research phase is aimed at developing the narrative and diving further into the selected data through additional analyses.

Third, they generate the necessary data and analyse them (in line with Duangphummet and Ruchikachorn's 2021 data preparation). Fourth, they interpret the results. For instance, the analyses of the occurrence of the terms representing the fake news discourse showed that three terms, nepberichten, fake berichten en nepinformatie, were hardly encountered in the data. This caused the researcher to not go further into these and focus on the other terms in his research.

Fifth, they create the narrative and tell the story in iterations with step three and four. For instance, the political scientist involved in the data story on the elections helped to interpret the consecutive analyses and results. Based on his expertise we interpreted the findings within the larger context of Dutch politics and concluded that the national public broadcaster failed to create a sufficiently level playing field for all parties participating in the elections. The presence of specific male ministers in relation to COVID-19, for instance in news programmes and press conferences, and the reuse of short clips from these, will probably have played into this.

[11] See `https://mediasuitedatastories.clariah.nl/tegenlicht-20-years`.

Again, these steps are iterative and going back and forth will be necessary to properly develop a sound Media Suite Data Story. Once the analyses have been completed and the story written, it is time to move on to the next phase: review.

5.3 Review

The goal of the review phase is to ensure both the technological quality of the Media Suite Data Story and the quality of its narrative and knowledge transfer. Fickers and Clavert (2021, par.29) point to the challenge of peer review and refer to it as the evaluation of "the results of a research, its methodology, its code and its data". Unfortunately, the authors do not elaborate on how they went about this and for now it remains "not fully solved". For Media Suite Data Stories, we invite experts not involved in the production and/or writing to review them. Domain experts review the story's content: its narrative and the conclusions based on the analyses. They focus on the data story's credibility and contribution to their discipline. They then deliver recommendations with respect to the research domain. For instance, the Media Suite Data Story on MediaOorlog was reviewed by an assistant professor in Media Studies whose research focuses on historical and archival media. Her expertise allowed her to assess whether the data story's narrative was credible and sufficiently transparent.

Technical reviewers focus on the data used for the research, the code, the analyses, and the connection with the underlying data. They deliver recommendations with respect to the technical aspects of the Media Suite Data Story. For instance, the Media Suite Data Story on *Tegenlicht* was reviewed by a colleague from the CLARIAH consortium. For other data stories. the data scientists involved also scrutinized each other's work.

In addition, we consider whether there are any legal aspects that need attention (such as potential copyright infringements or privacy issues).

The resulting recommendations are considered by the author(s) and implemented where deemed desirable and feasible. For example, one of the reviewers for the data story on the elections, a social and behavioural sciences scholar working on politics and democracy, also suggested including data from polls on the expected size of parties. However, due to limited resources, we decided not to follow up on this. He also suggested extending our methodological section, which we did.

Once this process is finished, the final phase starts: the Media Suite Data Story is published.

5.4 Publication

The publication phase is aimed not only at publication itself but also at checks, promotion and monitoring of the Media Suite Data Story. To this end, the Media Suite Data Story is published on the Media Suite Data Stories platform, after which the layout is checked for errors. Once approved, efforts are made to publicise the Media Suite Data Story through means and channels relevant to the research domains and related communities. This may include press releases, newsletters, blog posts, social media posts and other means. Finally, we monitor responses to the Media Suite Data Stories. For instance, the Media Suite Data Story on the elections was picked up by national public news broadcaster NOS and this sped up our publication process, so that NOS could write their own story on the findings.[12] We later finished the research as we

[12] See `https://nos.nl/artikel/2372814-onderzoek-rechts-domineert-de-verkiezingscampagne-op-radio-en-tv`.

intended, resulting in two versions.

At the time of writing, six Media Suite Data Stories by scholars have been published and we are working on another two. We use the protocol described above as a framework for the work and we intend to update it where necessary.

6 Conclusion and future work

In this article we have discussed the use of large data sets to create data stories in journalism and in academia. Although the process of creating data stories through exploration, analysis, visualisation, narration, and publication largely overlaps, journalists often use readily available data sets for visualisations, relying on pre-justified knowledge claims (Ettema and Glasser, 2006). In academia, new initiatives support the publication of data-driven research and data stories. These initiatives, of which the Media Suite Data Stories is one, explicitly include the exploration of data, the 'production' (D'Ignazio and Klein, 2020; Drucker, 2011; Schöch, 2013) of data sets and their analysis. Because the Media Suite provides access to such a varied and heterogeneous set of collections, its data are 'big' but messy. Exploring which collection might be most suitable to find an answer to a research question or hypothesis, thereby producing 'smart data' (Schöch, 2013), is part and parcel of the research and of making Media Suite Data Stories.

While the exploration phase focuses on data and serves to develop a research interest and related question or hypothesis in tandem with a suitable data set, the research phase is aimed at executing the research and developing the narrative. Visualisations play a major role in both, as a procedural tool (Stalph and Heravi, 2021), as they guide the understanding of the data and communicate findings with respect to the research question or hypothesis. The review phase should secure the quality of both the narrative and the data, tools and analyses underlying it, and publication includes efforts to raise awareness and monitor the Media Suite Data Story.

Our protocol presented above is aimed at securing the quality of this process and its outcomes. As Fickers and Clavert (2021) have acknowledged, the creation of these kinds of research narratives is a lot of work (see also Figueiras and Vizoso, 2022, with respect to journalistic data visualisations). We are continually learning and evaluating to find the best way forward.

One of the main challenges we face is making the creation of Media Suite Data Stories more appealing and accessible to academics. Researchers in the Humanities are still mainly trained in qualitative research methods and taking a quantitative approach is not a matter of course, as we experienced. To facilitate research with data available through the Media Suite, we are working on an environment that allows scholars to analyse Media Suite data using Jupyter Notebooks. This will not only facilitate research with the Media Suite data and metadata, but also support the critical scrutiny of research.

To further lower the threshold for academics, we are also working on three other ideas. First, we plan to make available a number of reusable notebooks. We aim to develop these as modules or building blocks for Media Suite Data Stories. By modules or building blocks, we mean specific queries, analyses and/or visualisations that can be adapted to researchers' needs without much knowledge of code. For instance, for a Media Suite Data Story on the discourse on 'fake news' in the Netherlands we created a query that specifically searches for the eight o'clock news broadcasts of the Dutch public broadcaster, a challenge given the many ways in which these programmes have

been archived in terms of metadata. Considering that future researchers might well be interested in including this selection in their research, we saved this query for future use. Researchers may reuse it with the option to change, for example, the period they wish to research. Another example is the analysis of the occurrence of a specific term, or a combination of terms, in this collection of newscasts resulting from the query. Researchers may, without much hassle, adapt the terms they want to count rather than write new code. Such building blocks also help to create a shared epistemological ground, as researchers reuse existing methods in the form of data collection and analysis. This approach aligns with Leon's suggestion to publish code (Leon, 2021), specifically in literate programming languages such as Jupyter Notebooks to not only make the process and steps in the analysis explicit, but also to streamline analysis and quality control procedures, especially when the source data are unavailable for sharing. One of our future aims is also to make the code for the analyses used in each Media Suite Data Storry available upon publication.

Second, we aim to recruit scholars experienced in data research and/or teams of researchers and data scientists. By recruiting researchers or research teams more experienced in quantitative and data research, we aim to make the production less resource-intensive for the editorial team and facilitate the creation of more stories. And third, one of the data stories in development concerns a pilot project in collaboration with an academic journal, to 'proof the concept' of academic Media Suite Data Stories. By exploring the possibility of a Media Suite Data Story as a proper academic article developed in collaboration with an academic journal, we hope to help develop multimedia academic storytelling and facilitate the creation of future academic Media Suite Data Stories.

In addition, we need to optimise our workflow, as we have not yet found the best software to support us in the interdisciplinary collaboration between (media) scholars and data scientists. Also, we aim to develop more visual styles of communication. Finally, future work also consists of developing a Data Suite, to make information about Media Suite data more readily available to potential authors.

7 Acknowledgements

We would like to extend our appreciation and thanks to three anonymous reviewers for their very useful feedback on a draft manuscript of this article. This work was made possible by the CLARIAHPLUS project funded by NWO (Grant 184.034.023).

References

Ángel Arrese. "In the Beginning Were the Data": Economic Journalism as/and Data Journalism. *Journalism Studies*, 23(4):487–505, March 2022. ISSN 1461-670X, 1469-9699. doi: 10.1080/1461670X.2022.2032803. URL https://www.tandfonline.com/doi/full/10.1080/1461670X.2022.2032803.

Alessia Bardi and Paolo Manghi. A Framework Supporting the Shift from Traditional Digital Publications to Enhanced Publications. *D-Lib Magazine*, 21(1/2), January 2015. ISSN 1082-9873. doi: 10.1045/january2015-bardi. URL http://www.dlib.org/dlib/january15/bardi/01bardi.html.

Alberto Cairo. *How charts lie: getting smarter about visual information*. W. W. Norton & Company, New York, first edition edition, 2019. ISBN 978-1-324-00156-0.

Jenna Condie and Cristina Costa. (Re-)exploring the practical and ethical contexts of digital research. In Cristina Costa and Jenna Condie, editors, *Doing Research In and On the Digital: Research Methods across Fields of Enquiry*, pages 205–212. Routledge, London, 1 edition, May 2018. ISBN 978-1-315-56162-2. doi: 10.4324/9781315561622. URL https://www.taylorfrancis.com/books/9781317201915.

Catherine D'Ignazio and Lauren F. Klein. *Data feminism*. Strong ideas series. The MIT Press, Cambridge, MA, 2020. ISBN 978-0-262-04400-4.

Johanna Drucker. Humanities Approaches to Graphical Display. *Digital Humanities Quarterly*, 005(1), March 2011. ISSN 1938-4122.

Apiwan Duangphummet and Puripant Ruchikachorn. Visual Data Story Protocol: Internal Communications from Domain Expertise to Narrative Visualization Implementation:. In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - IVAPP*, pages 240–247. SciTePess, 2021. ISBN 978-989-758-488-6. doi: 10.5220/0010327602400247. URL https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0010327602400247.

Faten El Outa, Matteo Francia, Patrick Marcel, Veronika Peralta, and Panos Vassiliadis. Towards a Conceptual Model for Data Narratives. In Gillian Dobbie, Ulrich Frank, Gerti Kappel, Stephen W. Liddle, and Heinrich C. Mayr, editors, *Conceptual Modeling, Proceedings 39th International ER Conference*, pages 261–270, Cham, 2020. Springer International Publishing. ISBN 978-3-030-62521-4 978-3-030-62522-1. doi: 10.1007/978-3-030-62522-1_19. URL http://link.springer.com/10.1007/978-3-030-62522-1_19. Series Title: Lecture Notes in Computer Science.

Faten El Outa, Patrick Marcel, Veronika Peralta, Raphaël da Silva, Marie Chagnoux, and Panos Vassiliadis. Data Narrative Crafting via a Comprehensive and Well-Founded Process. In Silvia Chiusano, Tania Cerquitelli, and Robert Wrembel, editors, *Advances in Databases and Information Systems*, volume 13389, pages 347–360, Cham, 2022. Springer International Publishing. ISBN 978-3-031-15739-4 978-3-031-15740-0. doi: 10.1007/978-3-031-15740-0_25. URL https://link.springer.com/10.1007/978-3-031-15740-0_25. Series Title: Lecture Notes in Computer Science.

James S. Ettema and Theodore L. Glasser. On the Epistemology Of Investigative Journalism. In G. Stuart Adam and Roy Peter Clark, editors, *Journalism: the democratic craft*, pages 126–140. Oxford University Press, New York, 2006. ISBN 978-0-19-518206-4 978-0-19-518207-1. OCLC: ocm60500442.

Andreas Fickers and Frédéric Clavert. On pyramids, prisms, and scalable reading. *Journal of Digital history*, (1), October 2021. URL https://journalofdigitalhistory.org/en/article/jXupS3QAeNgb. Publisher: DeGruyter Section: jdh001.

Ana Figueiras and Ángel Vizoso. Infomation Visualization: Features an Challenges in the Production of Data Stories. In Jorge Vázquez-Herrero, Alba Silva-Rodríguez, María-Cruz Negreira-Rey, Carlos Toural-Bran, and Xosé López García, editors, *Total Journalism: Models, Techniques and Challenges*, number volume 97 in Studies in Big Data, pages 83–96. Springer, Cham, Switzerland, 2022. ISBN 978-3-030-88028-6 978-3-030-88027-9.

Ben Fry. *Visualizing Data. Exploring and Explaining Data with the Processing Environment*. O'Reilly Media, Inc, Sebastopol, CA, 2008. ISBN 978-0-596-51455-6. OCLC: ocn190865378.

Jonathan Gray and Liliana Bounegru. Introduction. In Liliana Bounegru and Jonathan Gray, editors, *The Data Journalism Handbook: Towards A Critical Data Practice*, pages 11–23. Amsterdam University Press, Amsterdam, The Netherlands, March 2021. ISBN 978-90-485-4207-9 978-94-6298-951-1. doi: 10.5117/9789462989511. URL https://www.aup.nl/en/book/9789462989511.

Donna Haraway. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, 14(3):575–599, 1988. ISSN 00463663. doi: 10.2307/3178066. URL https://www.jstor.org/stable/3178066?origin=crossref.

Alexander Benjamin Howard. The Art and Science of Data-Driven Journalism. Technical report, Tow Centre for Digital Journalism, Columbia University, New York, NY, 2014. URL https://academiccommons.columbia.edu/doi/10.7916/D8Q531V1. Publisher: Columbia University.

Helen Kennedy and Martin Engebretsen. Introduction: The relationships between graphs, charts, maps and meanings, feelings, engagements. In Martin Engebretsen and Helen Kennedy, editors, *Data visualization in society*, pages 19–32. Amsterdam University Press, Amsterdam, 2020. ISBN 978-90-485-4313-7. OCLC: 1151404991.

Rob Kitchin. *Data lives: how data are made and shape our world*. Bristol University Press, Bristol, 2021. ISBN 978-1-5292-1564-9 978-1-5292-1514-4.

Sam Leon. Accounting for methods: Spreadsheets, scripts and programming notebooks. In Liliana Bounegru and Jonathan Gray, editors, *The Data Journalism Handbook*, pages 128–137. Amsterdam University Press, Amsterdam, The Netherlands, March 2021. ISBN 978-90-485-4207-9 978-94-6298-951-1. doi: 10.5117/9789462989511. URL https://www.aup.nl/en/book/9789462989511.

Wilson Lowrey, Ryan Broussard, and Lindsey A. Sherrill. Data journalism and blackboxed data sets. *Newspaper Research Journal*, 40(1):69–82, March 2019. ISSN 0739-5329, 2376-4791. doi: 10.1177/0739532918814451. URL http://journals.sagepub.com/doi/10.1177/0739532918814451.

Anthony McCosker and Rowan Wilken. Rethinking 'big data' as visual knowledge: the sublime and the diagrammatic in data visualisation. *Visual Studies*, 29(2):155–164, May 2014. ISSN 1472-586X, 1472-5878. doi: 10.1080/1472586X.2014.887268. URL http://www.tandfonline.com/doi/abs/10.1080/1472586X.2014.887268.

Liliana Melgar, Marijn Koolen, Hugo Huurdeman, and Jaap Blom. A Process Model of Scholarly Media Annotation. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval - CHIIR '17*, pages 305–308, Oslo, Norway, 2017. ACM Press. ISBN 978-1-4503-4677-1. doi: 10.1145/3020165.3022139. URL http://dl.acm.org/citation.cfm?doid=3020165.3022139.

Liliana Melgar-Estrada, Marijn Koolen, Kaspar Beelen, Hugo Huurdeman, Mari Wigham, Carlos Martinez-Ortiz, Jaap Blom, and Roeland Ordelman. The CLARIAH Media Suite: a Hybrid Approach to System Design in the Humanities. In *Proceedings*

of the 2019 Conference on Human Information Interaction and Retrieval, pages 373–377, Glasgow Scotland UK, March 2019. ACM. ISBN 978-1-4503-6025-8. doi: 10.1145/3295750.3298918. URL https://dl.acm.org/doi/10.1145/3295750.3298918.

Roeland Ordelman, Liliana Melgar, Carlos Martinez-Ortiz, and Julia Noordegraaf. Media Suite: Unlocking Archives for Mixed Media Scholarly Research. In Inguna Skadina and Maria Eskevich, editors, *CLARIN Annual Conference 2018 Proceedings*, pages 21–25, Pisa, Italy, October 2018. CLARIN. URL https://ris.utwente.nl/ws/portalfiles/portal/63914792/CE_2018_1292_CLARIN2018_ConferenceProceedings.pdf#page=28.

Sylvain Parasie. Data-Driven Revelation?: Epistemological tensions in investigative journalism in the age of "big data". *Digital Journalism*, 3(3):364–380, May 2015. ISSN 2167-0811, 2167-082X. doi: 10.1080/21670811.2014.976408. URL http://www.tandfonline.com/doi/full/10.1080/21670811.2014.976408.

Nathalie Henry Riche, Christophe Hurter, Nicholas Diakopoulos, and Sheelagh Carpendale. Introduction. In Nathalie Henry Riche, Christophe Hurter, Nicholas Diakopoulos, and Sheelagh Carpendale, editors, *Data-driven storytelling*, A K Peters Visualization Series, pages 1–15. CRC Press/Taylor & Francis Group, Boca Raton, FL, 2018. ISBN 978-1-138-19710-7 978-1-138-48225-8.

Lars Rinsdorf and Raoul Boers. The need to reflect. Data journalism as an aspect of disrupted practice in digital journalism and in journalism education. In *Proceedings of the International Association for Statistical Education.*, Berlin, 2016. IASE. ISBN 978-90-73592-37-7. URL https://iase-web.org/documents/papers/rt2016/Rinsdorf.pdf.

Christof Schöch. Big? Smart? Clean? Messy? Data in the Humanities. *Journal of Digital Humanities*, 2(3), 2013. URL http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/.

E. Segel and J. Heer. Narrative Visualization: Telling Stories with Data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1139–1148, November 2010. ISSN 1077-2626. doi: 10.1109/TVCG.2010.179. URL http://ieeexplore.ieee.org/document/5613452/.

Florian Stalph and Bahareh Heravi. Exploring Data Visualisations: An Analytical Framework Based on Dimensional Components of Data Artefacts in Journalism. *Digital Journalism*, pages 1–23, August 2021. ISSN 2167-0811, 2167-082X. doi: 10.1080/21670811.2021.1957965. URL https://www.tandfonline.com/doi/full/10.1080/21670811.2021.1957965.

Charles D. Stolper, Bongshin Lee, Nathalie Henry Riche, and John Stasko. Emerging and Recurring Data-Driven Storytelling Techniques: Analysis of a Curated Collection of Recent Stories. April 2016. URL https://www.microsoft.com/en-us/research/publication/emerging-and-recurring-data-driven-storytelling-techniques-analysis-of-a-curated-collection-of-recent-stories/.

Jasmijn Van Gorp, Liliana Melgar Estrada, and Julia Noordegraaf. Involving Users in Infrastructure Development: Methodological Reflections From the Research Pilot Projects Using the CLARIAH Media Suite. *TMG Journal for Media History*, 24(1-2), December 2021. ISSN 2213-7653, 1387-649X. doi: 10.18146/tmg.809. URL https://www.tmgonline.nl/article/10.18146/tmg.809/.

Wibke Weber. Exploring narrativity in data visualization in journalism. In Martin Engebretsen and Helen Kennedy, editors, *Data Visualization in Society*, pages 295–311. Amsterdam University Press, Amsterdam, April 2020. ISBN 978-94-6372-290-2 978-90-485-4313-7. doi: 10.5117/9789463722902. URL https://www.aup.nl/en/book/9789463722902.

Wibke Weber, Martin Engebretsen, and Helen Kennedy. Data stories. Rethinking journalistic storytelling in the context of data journalism. *Studies in Communication Sciences*, 18(1):191–206, November 2018. ISSN 2296-4150, 1424-4896. doi: 10.24434/j.scoms.2018.01.013. URL https://www.hope.uzh.ch/scoms/article/view/j.scoms.2018.01.013.

Florian Windhager, Paolo Federico, Günther Schreder, Katrin Glinka, Marian Dörk, Silvia Miksch, and Eva Mayr. Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges. *IEEE Transactions on Visualization and Computer Graphics*, 25(6):2311–2330, June 2019. ISSN 1941-0506. doi: 10.1109/TVCG.2018.2830759. Conference Name: IEEE Transactions on Visualization and Computer Graphics.