Managing a “plurality of desires”: different support strategies for scholars and journalists in the building of a Data Stories infrastructure

Willemien Sanders (Utrecht University, Sound & Vision), Mari Wigham (Sound & Vision), Rana Klein (Sound & Vision), Roeland Ordelman (Sound & Vision), Jasmijn Van Gorp (Utrecht University) and Julia Noordegraaf (University of Amsterdam)

Work in progress - please do not cite without permission from the authors.

Abstract

The Clariah Media Suite provides online access to a wide array of audio-visual archival collections and has been developed with and for scholars in a range of digital humanities domains, from media studies to history. Currently, we are developing an infrastructure for presenting research with Media Suite data and metadata: Media Suite Data Stories (see Sanders et al. 2022). Professional users, such as journalists, are also showing an interest in the Media Suite and related infrastructure.

, However, there seems to be little discussion of how the practices and demands between such different users vary, to which Guldi (2020, par.77) refers as a “plurality of desires”, or of the consequences of this plurality for data providers and editorial teams. The variations in their requirements and the contexts in which they work need to be taken into account when designing new and improved features and functionalities for the Media Suite. In addition, the Data Stories editorial team needs to develop different strategies for supporting these users in terms of access, tools, and research support.

We discuss five aspects in which scholars and journalists require different kinds and levels of support: speed, transparency, reflection, data/service provider involvement and sensitivity/copyright issues. We use three cases to discuss these aspects: an academic researcher investigating digitized historical newspapers, a journalist using a pre-set analysis of contemporary news and current affairs programmes to examine a specific discourse, and an investigative journalist probing the language use of a controversial Dutch public broadcaster.

Based on an in-depth discussion of these use cases, we argue that the Media Suite Data Stories editorial team needs to balance the support of the demands of different users with its own resources, legal, and technical possibilities .

Keywords

Clariah Media Suite, journalism, digital methods, data stories, critical humanities

Introduction

The Clariah Media Suite provides online access to a wide array of audio-visual archival collections and has been developed with and for scholars in a range of digital humanities domains, from media studies to history. Currently, we are developing an infrastructure for presenting research with Media Suite data and metadata: Media Suite Data Stories (see Sanders et al. 2022). This infrastructure is hosted by the Netherlands Institute for Sound and

Vision, in the role of data provider cum service provider. The production of Media Suite Data Stories is supported by an editorial team, the authors of this paper.

The use of Media Suite data and Media Suite Data Stories have proven be to interesting for both scholars and journalists (see the current list of Media Suite Data Stories at https://mediasuitedatastories.clariah.nl (in Dutch)). Scholars acknowledge the value of academic online collections and archives for both scholars and journalists (as well as, for instance, educators and the public at large, see Schafer and Winters 2021; Buddenbohm et al. 2021). However, as these two groups of professionals have very different demands, we as Data Stories editorial team need to adapt our support practices to meet these different demands. The development of these strategies is the focus of this paper. In this paper we discuss five aspects of research that scholars and journalists approach differently and what these differences mean for the support the Data Stories editorial team can provide for them. These aspects are: speed, transparency, reflection, data/service provider involvement and sensitivity/copyright issues.

These differences have instigated the development of three approaches to cater for scholars and journalists: individual support for (teams of) scholars, a template supporting discursive analysis of a set of news and current affairs television programmes that can be used by either scholars or journalists, and provisions for data access for independent publications by public broadcaster journalists. We discuss each in relation to a specific project.

We conclude that cultural heritage institutions need to acknowledge and understand the different needs and capabilities of different users, including scholars and journalists, in order to match their own resources to the needs of these users. First, we will briefly introduce Media Suite Data Stories in more detail.

About Media Suite Data Stories

Media Suite Data Stories are stories based on research with Media Suite data and specifically (enriched) metadata, such as subtitles, automatic speech recognition files (ASR), and data about individuals included in the Dutch common thesaurus for audio-visual archives, the Gemeenschappelijke Thesaurus Audiovisuele Archieven (GTAA), for instance through the recognition of their names, voices and/or faces. It provides access to a large body of work that represents a historical public discourse, albeit subject to the shortcomings of that discourse.

Media Suite Data Stories are narratives that are driven by the interpreted results of quantitative Media Suite (meta)data analyses. Their production is supported by the editorial team. The creation of Data Stories requires special attention as it demands a variety of expertise: the authors need to have knowledge of data, tools and the domain they are investigating, in order to draw accurate, well-founded conclusions. This makes the production of Data Stories by definition an interdisciplinary effort.

As Schafer and Winters (2021) observe, online collections and archives exist in ever changing social, political, technical, and legal contexts. While the Clariah Media Suite was developed for and with media scholars, scholars outside media studies are also showing interest in the Media Suite and in Data Stories, as are journalists (further discussed below). However, scholars and journalists work in vastly different professional cultures, despite similarities in their practices (see, for instance, Fry 2008). While their methods may be similar when it comes to the activities of gathering, processing, and analysing data and reporting the results, their work differs with respect to, amongst others, the following five aspects: speed, transparency, reflection, their relation to data and service providers and sensitivity and copyright issues. Below we will discuss each in more detail, before addressing what this means for the way the Data Stories editorial team can support these different authors.

Five aspects of research practices

The first three aspects we discuss relate to the users themselves, the scholars and journalists, while the other two relate to the data providers.

Speed

Generally, the speed with which new content needs to be delivered to audiences is much higher for journalists than for scholars. This 'need for speed' in journalism is often attributed to commercial demands and increasing competition, but to bring 'the news' as quickly as possible is also at the heart of journalism (Juntunen 2010). Juntunen (2010) argues that speed does not necessarily come at the expense of other professional and ethical considerations, such as accountability, and should be considered as one of a number of journalistic values. Scholarship, by contrast, is often slow, and going from a first draft of an article or chapter to publication often takes over a year.

The internet and, subsequently, social media have increased the speed at which information travels and journalists have become part of this online infrastructure (Phillips 2012). In academia, while speed has increased due to technological developments, it is not necessarily an academic value.[1] Arguably, investigative journalists assigned to an in-depth project have more time than daily reporters, and might be positioned between reporters and scholars when it comes to speed: although there is more time for an in depth understanding of data and their context (see D'Ignazio and Klein 2020 for a discussion of the importance of context), current events still affect the decision to pursue a story and to publish it earlier or later (see, for instance, Sanders 2020).

The Media Suite Data Stories editorial team has limited resources. With respect to creating Media Suite Data Stories this means that the editorial team needs to consider how it might best employ its limited resources to facilitate the creation of Media Suite data for scholars and journalists in ways that acknowledges their different 'speeds'.

Transparency

Although science has long tried to maintain an aura of objectivity, with the researcher acting as a disinterested party who merely registers or collects data for analysis, contemporary approaches acknowledge and further develop Haraway's (1988) idea of the 'situatedness' of researchers. Scholars nowadays have an obligation to be transparent and account for their methods to such an extent that other scholars may not only replicate the research but also

[1] Recently, for instance, speed did matter in the search for a vaccine against the corona virus.

understand how the professional and personal background of the author(s) may have informed their choices, decisions, and interpretations.

According to Curry and Stroud (2021), a similar development has taken place in journalism, where the unattainable goal of ‘objective’ reporting has also been substituted by a drive for transparency, especially after the introduction of the internet. Curry and Stroud define transparency as “a news organization’s openness about its journalistic practices and decision-making processes” (903). Domingo and Heikkilä (2012: 272) distinguish between ‘actor transparency’, which relates to journalists’ professional affiliations, and ‘production transparency’, which relates to “how they gather and select the news, and how they deal with their sources”. The latter in particular is relevant for the current discussion. Various scholars have discussed transparency in journalism as an “antidote” to the declining trust in journalists (Curry and Stroud 2021: 902), which makes it more relevant than ever.[2]

For both scholars and journalists, transparency is important for the credibility of their work. The main difference resides in two aspects: replicability versus understanding and the scholarly audience versus the general audience.

Scholars need to facilitate the replication of their research in order for colleagues to understand the process as well as extend the research to other areas using the same methodology. This way, scholars can build on each other’s work and develop related knowledge. For journalists, and especially investigative journalists, the main responsibility is to account for their work towards the general audience, so that they understand the basis for any claims about the real world that journalists may make. While scholars can use their own discipline specific jargon to do so, journalists need to translate discourse from specific fields to more colloquial language to make the information accessible for people with different literacies. Although, arguably, this need to facilitate different literacies has also been acknowledged by academia, scholars in general still rely on a lot of jargon.

Reflection

A third aspect, related to both speed and transparency, is the variation in the degree of reflection on the methods used. With reflection we mean a careful and critical (re)consideration of the methods used and their limitations, and what this means for the outcomes, as well as a proper effort to understand the relationship between data, analysis tools (calculations, AI) and results. A profound reflection on the methods will help develop accurate and nuanced conclusions, whereas a lack of reflection runs the risk of drawing overly quick and inaccurate conclusions. In particular, Data Stories require reflection on the data and tools used, i.e., data criticism and tool criticism, as historical data sets include, for instance, data breaches as a consequence of gaps and changes in metadata policies and practices. Metadata enrichments such as ASR and subtitles may contain errors, for instance due to the use of foreign language. Users of any kind who are new to the use of tools and data may not grasp the need to critically examine data and tools just as rigorously as they would do other sources and methods.

[2] At the same time, both journalists and scholars may actively conceal their sources in certain circumstances, such as where the identification of a human source could put that person in danger.

Although arguably no single humanities research project is fully linear, in data research in particular the process is characterised by iterations: initial analyses and their interpretations will inform subsequent analyses and may also lead to refinement or even alteration of the research question. Scholars can only be transparent about their methods if they critically reflect on this process. They should therefore take the time to consider the analysis results in the context of their data set and its characteristics, and in the context of the analysis tools they used (see, for instance Koolen, van Gorp, and van Ossenbruggen 2019).

Ideally, journalists take the same approach, However, according to D'Ignazio and Klein (2020), journalists often work with existing data sets, offered, for instance, by government bodies and NGO's. Such data sets have been structured and 'cleaned up' to make them ready for use. They represent "bureaucratic accounts" and thereby offer "pre-justified accounts" of social realities (Ettema and Glasser 2006: 129). Such practices will make critical reflection less of a matter of course for journalists than it is for scholars.

Data/service provider involvement

Fourth, the involvement of the data and/or service provider (typically archives and libraries) is different in projects involving scholars than in projects involving journalists. There is often a common interest between data providers, such as cultural heritage organizations, and scholars, which can lead to various forms of active cooperation and facilitation. The data provider supports the academic's understanding of the data, their analysis and reflection on the process.

As discussed above, journalists often work with existing data setsAccording to D'Ignazio and Klein (2020) they often lack contextual information about the purpose and methods used for data collection, further impeding critical reflection. The data available via the Media Suite is much more messy. Due to differences in metadata and collection practices over time, datasets from one era may be hard to compare with data sets from other eras. Therefore, journalists need to learn how to understand these data and what its characteristics might mean for the interpretation of the results. This requires expertise from the data provider, which must be obtained in some other manner if active cooperation is not possible.

Sensitivity/copyrights issues

Finally, there is the issue of sensitive and/or copyright protected data, which has consequences for the means and level of access to these data, and the measures which must be taken to prevent the data leaking. The data provided via the Media Suite are to a large extent copyright protected and there are license agreements with the owners of the data that specify who may access them. Given the nature of the Media Suite as an academic resource, all those affiliated with Dutch higher education institutions can access the Media Suite data, through an institutional login.

As mentioned above, the data offered through the Media Suite also has value for journalists. However, journalists only have access to the metadata, per individual item via the user interface. If they want access to the data (including streamed video content, images, audio content newspaper scans), or to metadata for a large number of items, such as transcripts for a TV series, then they need to make individual arrangements with the Data

Stories editorial team. Even then, copyright considerations may render it impossible for the data to be supplied to them.

Managing academic and journalistic pluralities of desire

In the past two years, the Data Stories editorial team has developed data stories with both scholars and journalists. This has led to the continued development of the infrastructure for creating and publishing Media Suite Data Stories. It has also led to the following insights and solutions, which we see by no means as definitive or exhaustive, but rather as a step towards further development: individual support for (teams of) scholars; a template for discursive analysis for scholars and journalists; and provisions for access to Sound & Vision data for public broadcaster journalists.

Teams of scholars and data scientists

The Media Suite Data Stories editorial team has been involved in developing and writing a number of Data Stories. As Fickers and Clavert (2021) justly state, doing data research and creating narratives that report such research is a lot of work. It was very useful for the team to experience this work themselves. It allowed them to gain a deeper insight into user requirements, to more clearly identify the foundations of a good data story, and to develop a model workflow for creating a well-founded data story. The data stories created also provide examples and inspiration to scholars. However, the team does not have the resources to continue to be involved in the development of Media Suite Data Stories at this level. Nor is this desirable, as the added value of the editorial team lies in facilitating the production of Media Suite Data Stories, not conducting novel data research.

Instead, the infrastructure and knowledge built up by the team can be used by scholars to create their own data stories. The editorial team provides support in terms of logistics and advice, but it is up to the scholar to execute the research and report the results. The need for transparency demands that scholars are sufficiently capable of understanding their methods and reporting about their research processes. Therefore, scholars who wish to create a Media Suite Data Story need to either be sufficiently experienced in doing data research or work together with data science colleagues. Such scholars or teams must be sufficiently capable of reflecting on their research as well. The editorial team can support transparency and reflection by ensuring that the scholarly team is aware of their necessity.

While there are many projects involving collaboration between data providers and scholars, this is not currently the case for journalists. As journalists are also limited in their use of Media Suite infrastructure, this solution is best suited to scholars.

This solution aids transparency and reflection by employing multi-disciplinary teams. As this solution is only for scholars, speed is less of an issue, as is copyright.

A template for discursive analysis

So far, research for Media Suite Data Stories has often used the television archive of Sound & Vision, as television is one of the major mass media that report on current affairs. For such research, either the subtitles (produced for hearing impaired audiences) or ASR files have been used. In such files it is possible to count, for instance, the number of times certain terms appear in speech. This method was used for the Data Story about the 2021 Dutch parliamentary elections as well as the Data Story about fake news discourse in the Netherlands.[3]

To facilitate this kind of research, the editorial team has created a template, based on the latter of these two stories. The template facilitates the search for a number of terms, chosen by the author, in a corpus of programmes with subtitles and speech transcripts, created by a fixed query for a selection of television news and current affairs programmes over a fixed period of time. For these terms, it provides statistics on their occurrences in programmes, over time, per genre and per broadcaster. It also provides statistics for the entire corpus, to allow the author to assess its balance and completeness. It is up to the author to make further selections if necessary (for instance, to discard data outside of the period of interest) and to create visualisations to further explore and understand the data and results.

This template makes it possible to tell a wide variety of stories about the public television discourse, without requiring that the author have a data science background or forcing them to spend a lot of time constructing a corpus. It does demand that authors thoroughly think through the questions they wish to answer and the terms they wish to investigate in advance, while requiring a minimum amount of support from the editorial team. The template is flexible in the sense that the search terms and the data set may easily be adapted to the author's needs.

The template answers to some degree the 'need for speed' as it provides a low threshold way of investigating a certain popular discourse. It aids transparency as the framework for and steps within the research are pre-set. The built-in dataset checks help raise awareness of the need for a balanced and sufficiently complete dataset. Also, it supports the development of an in-depth knowledge and understanding of the data set over time, as different authors use and interpret it from different perspectives, resulting in a collaborative and cumulative reflection. The involvement of the editorial team is minimal. As the template supplies statistical results, and not individual programme metadata or transcripts, sensitive and copyrighted material is protected.

The template is currently being tested for a new data story on refugees. Hopefully this template holds promise for both scholars and journalists.

Access for public broadcaster journalists

The rights to a large part of the television and radio archive hosted by Sound & Vision are owned by the Dutch public broadcasting umbrella organisation NPO. Journalists working for NPO programmes already have access to the archive through another Sound & Vision service, aimed at finding and ordering material for reuse, not analysis. However, because they are not affiliated with an academic institution, they are not entitled to full access to the data and enriched metadata in the Media Suite and related infrastructure.

Recently, a journalist working for Pointer, an NPO digital platform for investigative journalism, contacted the Media Suite Data Stories editorial team with a wish to investigate metadata in the form of the subtitles of the programmes of a specific public broadcaster. In consultation with the legal department at Sound & Vision, a provisional solution was found

[3] See https://mediasuitedatastories.clariah.nl/elections-dec-2021 and https://mediasuitedatastories.clariah.nl/fake-news-2023 (both in Dutch).

by providing in-house supervised access. Only the (aggregated) results of analyses could be exported for further exploration.

This solution expedited access within the timeframe of the journalistic investigation while protecting copyrighted material. Transparency on the production of the dataset itself was provided. Further transparency and reflection were the responsibility of the journalists, as the editorial team was not involved in the analysis process nor in the interpretation of results. For journalists who lack the expertise or support network to meet these responsibilities, this is therefore not a viable solution.

While this collaboration underlines the relevance of the metadata in the Media Suite for journalism, providing access in this way is not a scalable or sustainable solution as it required the editorial team to compile the necessary data, check it for copyright issues, and then supervise the journalist while they work with it. Sound & Vision, as a data and service provider, does not formally have the authority to allow journalists access to these data via the Media Suite. As a consequence, if journalists are considered to be relevant users of the (enriched) metadata in the Media Suite, the terms under which users may access these data should be reconsidered and renegotiated with the rightsholders.

Conclusion

In this paper we have discussed different ways in which scholars and journalists operate while conducting research and what this means for the support the Media Suite Data Stories editorial team may provide. We focused on five aspects: speed, transparency, reflection, data/service provider, and sensitivity/copyrights.

In facilitating individual scholars and scholarly teams in creating data stories, we support transparency and reflection by requiring that they possess the necessary data research skills. Developing Media Suite Data Stories requires iterations between analyses and interpretations and to do so responsibly, authors need to be able to reflect on the data and their context, methods and tools and results, and the relationships between them. We have developed a model workflow to support scholars in this process. The lack of formal collaborations between data providers and journalists, and the limitations on use of the Media Suite by journalists, make this solution best suited to scholars.

In order to adjust from the slow pace of in-depth research and reflection that scholars are used to, to the different ‘speed’ of journalists, we developed a low threshold template for the discursive analysis of the public television discourse. Such a template is also useful for scholars without data research skills. Using a fixed corpus query and offering a single data set based on the search for a predefined set of terms allows for a relatively quick understanding of the discourse surrounding those terms. Correct methods of calculating statistical distributions and knowledge of good practice, such as the importance of a balanced dataset, are built into the template. Transparency and collaborative and cumulative reflection are supported through the re-use of the data from different perspectives.

To give public broadcaster journalists more independence and flexibility in their research while still protecting sensitive and copyrighted data, we developed a solution of supervised access to curated datasets in-house. This solved the issue of data access for non-scholarly public broadcaster journalists who do not fall under the terms of use of the Media Suite. However, it required a lot of custom work, and is therefore not scalable or sustainable.

Our experience shows that the role of the Data Stories editorial team varies between different collaborations. The use of the Media Suite and related infrastructure by journalists forces the main data provider, Sound & Vision, to reassess its position, specifically when copyright issues come into play. If the archive wishes to facilitate journalists in their use of Sound & Vision data for data research in a sustainable and scalable manner, it should look into a solution for copyright issues, for example by attempting to renegotiate the terms of access with copyright owners.

To support the development of Media Suite Data Stories by different users, the editorial team needs strategies to ensure that each group has the access, tools and research support that suit them. This has resulted in different ways to facilitate the development of Media Suite Data Stories. In all three solutions discussed in this paper, the need for data criticism requires the scholar or journalist to be informed about how the data was produced. For this reason, the editorial team is also working on documentation of metadata and the underlying processes.

This work was made possible by the CLARIAH-PLUS project funded by NWO (Grant 184.034.023).

References

Buddenbohm, Stefan, Maaike de Jong, Jean-Luc Minel, and Yoann Moranville. 'Find Research Data Repositories for the Humanities - the Data Deposit Recommendation Service'. *International Journal of Digital Humanities* 1, no. 3 (July 2021): 343–62.

Curry, Alexander L, and Natalie Jomini Stroud. 2021. 'The Effects of Journalistic Transparency on Credibility Assessments and Engagement Intentions'. *Journalism* 22 (4): 901–18. https://doi.org/10.1177/1464884919850387.

D'Ignazio, Catherine, and Lauren F. Klein. 2020. *Data Feminism*. Cambridge, MA: The MIT Press.

Domingo, David, and Heikki Heikkilä. 2012. 'Media Accountability Practices in Online News Media'. In *The Handbook of Global Online Journalism*, edited by Eugenia Siapera and Andreas Veglis, 1st ed., 272–89. Wiley. https://doi.org/10.1002/9781118313978.ch15.

Ettema, James S., and Theodore L. Glasser. 2006. 'On the Epistemology Of Investigative Journalism'. In *Journalism: The Democratic Craft*, edited by G. Stuart Adam and Roy Peter Clark, 126–40. New York: Oxford University Press.

Fickers, Andreas, and Frédéric Clavert. 2021. 'On Pyramids, Prisms, and Scalable Reading'. *Journal of Digital History*, no. 1 (October). https://journalofdigitalhistory.org/en/article/jXupS3QAeNgb.

Fry, Ben. 2008. *Visualizing Data. Exploring and Explaining Data with the Processing Environment*. Sebastopol, CA: O'Reilly Media, Inc.

Guldi, Jo. 'Scholarly Infrastructure as Critical Argument: Nine Principles in a Preliminary Survey of the Bibliographic and Critical Values Expressed by Scholarly Web-Portals for Visualizing Data'. *Digital Humanities Quarterly* 14, no. 3 (1 September 2020). http://digitalhumanities.org:8081/dhq/vol/14/3/000463/000463.html.

Haraway, Donna. 1988. ‘Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective’. *Feminist Studies* 14 (3): 575–99. https://doi.org/10.2307/3178066.

Juntunen, Laura. 2010. ‘Explaining the Need for Speed. Speed and Competition as Challenges to Journalism Ethics’. In *The Rise of 24-Hour News Television: Global Perspectives*, edited by Stephen Cushion and Justin Lewis, 167–80. New York: Peter Lang.

Koolen, Marijn, Jasmijn van Gorp, and Jacco van Ossenbruggen. 2019. ‘Toward a Model for Digital Tool Criticism: Reflection as Integrative Practice’. *Digital Scholarship in the Humanities* 34 (2): 368–85. https://doi.org/10.1093/llc/fqy048.

Phillips, Angela. 2012. ‘Sociability, Speed and Quality in the Changing News Environment’. *Journalism Practice* 6 (5–6): 669–79. https://doi.org/10.1080/17512786.2012.689476.

Sanders, Willemien. 2020. *‘The Story That Could Have Been. Testing the Added Value of the CLARIAH Media Suite for Investigative Journalism’*. Hilversum, the Netherlands: Netherlands Institute for Sound and Vision.

Sanders, Willemien, Ordelman, Roeland, Wigham, Mari, Klein, Rana, Van Gorp, Jasmijn, & Noordegraaf, Julia. (2022, May 29). Developing Data Stories as Enhanced Publications in Digital Humanities. DH Benelux 2022 - ReMIX: Creation and alteration in DH (hybrid), Belval Campus, Esch-sur-, Luxembourg. https://doi.org/10.5281/zenodo.7501190.

Schafer, Valérie, and Jane Winters. ‘The Values of Web Archives’. *International Journal of Digital Humanities* 2, no. 1–3 (November 2021): 129–44.