DHQ: Digital Humanities Quarterly
2020
Volume 14 Number 3

Fading Away... The challenge of sustainability in digital studies

Christine Barats <c_dot_barats_at_wanadoo_dot_fr>, Cerlis, University of Paris Descartes
Valérie Schafer <Valerie_dot_schafer_at_uni_dot_lu>, C²DH, University of Luxembourg
Andreas Fickers <andreas_dot_fickers_at_uni_dot_lu>, C²DH, University of Luxembourg

> The dominant economy of the traces of each era are based on agents of different ages. Within the framework of slow-moving institutions whose role is precisely to make the system stagnate, the rules of this economy change at the faster pace of social changes, which it adjusts with technologies that change even faster - and which can produce unexpected or contrary effects on the logic of the first two levels. This stratification of collective equilibriums induces an interweaving of technical environments and innovative societal projects with older or even archaic beliefs or hierarchies. (Merzeau 1998, our translation) 1

Introduction

The question of future access to knowledge production was asked long before the digital era. The risk of losing (analogue) data dates back to before digitisation and care was reinforced for example through transfer of data onto magnetic strips before computer memory storage [Borghoff et al. 2006]. However, data preservation has recently attracted new attention within research: questions of sharing data and reproducibility of science, open access and maintenance have become more and more pertinent as the websites and digital productions come of age. “404 not found” messages now replace a growing number of hyperlinks, and the Web as scientific platform is full of digital wastelands, caused by the end of research projects. As underlined by Smithies and al., “finding a comprehensive and scalable approach to sustainable development in digital humanities labs is a non-trivial problem” [Smithies et al. 2019, 2] 2

In 2011, the Manifesto for Digital Humanities has made a plea “for open access to data and metadata, which must be documented and interoperable, both technically and conceptually” [Dacos 2012]. Transdisciplinary studies published around the same time underline that “sustainability challenges require new ways of knowledge production,” recommending a collaborative approach and new scientific agencies in the academic field [Lang et al. 2012]. Without doubt, the challenge of sustainability has turned into a key concern for the field of digital humanities at large as illustrated by the theme choosen by the Digital matters (an interdisciplinary research agenda) at the University of Utah.[1] Large scale research infrastructures such as Huma-Num, Dariah and OPERAS[2] seriously consider the sustainability of humanities and social sciences (HSS) research, and funding bodies increasingly make “data management plans” a must in research applications. 3

As defined by the Stanford Libraries: “A data management plan (DMP) is a written document that describes the data you expect to acquire or generate during the course of a research project, how you will manage, describe, analyse, and store those data, and what mechanisms you will use at the end of your project to share and preserve your data.”[3] These sentences underline that data management is not just about the data that are used, but also about those, which are produced (metadata, results of research, etc.). In fact, data management covers the whole life cycle of a research project, from data selection to their curation and description, from analyse and interpreting data to its publication and long term storage. Therefore, a sustainable stewardship of research data includes a strategy for the readability, maintenance and transparency of those data [Hoekstra and Koolen 2018]. While this plea for data stewardship – often based on the FAIR principles (FAIR stands for Findable, Accessible, Interoperable, and Reusable, see Wilkinson et al. 4

2016 and Mons 2018) – has become a shared ambition in earth or climate sciences, it remains a theoretical discussion rather than a shared practice in the field of humanities and social sciences.

This paper will focus in particular on HSS, in order to uncover the centrality of this question and the diversity of the researchers' responses to the many challenges at stake. The aim is to give a voice to researchers from different disciplines that are basically facing the same problems when it comes to data management issues and the question of sustainability. The present study examines all stages of research, from data collection to the dissemination of knowledge, and underlines that challenges are not limited to the technical maintenance of software, tools and data, but also apply to the wider institutional contexts, epistemic traditions and social practices in which the doing of research in humanities and social sciences are embedded. As highlighted by Bachimont (2017), a loss of intelligibility - that is an understanding of the context in which data are generated and interpreted – is as critical as the technical or material loss of data, if not more. Therefore, the topic at stake necessitates a broader reflection on the institutional, economic and political dimensions of knowledge production, dissemination and conservation in the digital era [Musiani and Schafer 2017]. 5

In order to identify the conflicts between the ephemeral, the flow of data, and the safeguarding of knowledge production for future needs, our analysis is largely based on personal experiences of French researchers in the field of digital studies. A qualitative survey was submitted to senior colleagues in the field who have carried out larger research projects and as such responsible for maintaining and preserving data. The survey includes colleagues working in disciplines such as history, information and communication studies, digital humanities, and linguistics. In total, twenty colleagues answered our indicative interview-based online survey[4] in the period between May 2018 and October 2018. The survey included five open questions concerning data access, methods and tools of data gathering and curation, approaches to data preservation, reflections on sustainability, and the place devoted to long-term preservation. Longer face-to-face or telephone interviews were carried out with ten respondents in order to dig deeper into specific questions or problems raised in the survey. 6

The aim of our project was not to produce strict guidelines for sustainable data management for open science in the field of social sciences or humanities. We fully agree with Smithies et al. (2019) that complex projects may share common and generic solutions, but also often need to rely on *ad hoc* solutions, taking into account the specific characteristics of a research project. We therefore want to emphasise the need to think about sustainability as a key element of digital hermeneutics, encompassing the critical reflection on the epistemic role of data management and stewardship in the entire life cycle of knowledge production and dissemination [Fickers 2020]. In promoting this holistic approach, we aim at addressing the inherent tension between different temporalities at stake: between the long-term needs of data preservation and maintenance on the one side, and the paradoxically short life cycles of the data formats, platforms and infrastructures on the other side. We seek to explore these tensions between the necessary stability and reproducibility [O' Sullivan 2019] of research and the instability of the digital domain it depends on at several levels and temporalities. The often-contradictory logics of these different temporalities will be discussed in a multilevel perspective, including the institutional, human and technical dimensions of a phenomenon we qualify as "fading away." This study therefore considers three key stages in research – namely data access and building a corpus, establishing a research framework and analysis, and finally the use/dissemination of results. 7

1. From data access to corpus building

"There are a great number of difficulties. The API of Facebook, a semi-private, semi-public social network, prevents automatic data crawling and makes the archiving of data a tedious process that involves numerous screenshots and the difficulty of organising manually collected data. I also encountered difficulties linked to the plasticity of devices and therefore the loss of data: the closure of pages or Facebook profiles, deleted comments or posts... as well as problems related to ethical issues for data collection: what data can I use, and how? Does access to 'private' data also permit the use of said data? Is the consent of the respondents sufficient to publish the data?" (ICS interview, July 2018). 8

By explaining her working methods, this scholar in information and communication studies highlights realities that are frequently invisible in research papers and remain hidden from public view: digital studies are more often than is usually 9 assumed a mixture of science and handicraft, a kind of “thinkering” [Fickers 2017] process which involves several challenges and difficulties along the way, and whose solutions may sometimes be very manual, empirical and pragmatic. This is far from the methodology which may then be outlined in a published paper, but this hidden struggling and the iterative logic of science in the making remains a reality for many research projects that are characterised by a mixed method and hands-on learning by doing approach trying to deal with screenshots, time-consuming manual curation, changing access conditions, obsolete hyperlinks, unclear legislation – in short, the normal grappling with the instability of data, devices, and tools.

1.1 Diversity and instability of data search and gathering

10 Researchers using digital data in the field of humanities and social sciences mainly work with social media data originating from commercial platforms such as Facebook, Twitter or Instagram. These data are often collected in real time. In the following paragraph, we will discuss some French examples of such projects in order to highlight the different possibilities of scraping or gathering online data and to reflect on both the opportunities and challenges that come with the different approaches.

11 Twitter-data are certainly one of the favourite digital-born sources to be studied by digital humanists. French examples are the study carried out by Badouard (2016) on the #jenesuispasCharlie hashtag after the attacks against the editorial staff of Charlie Hebdo in January 2015, and Frédéric Clavert’s project on the commemorations of the First World War [Clavert 2018]. As noticed by Dagiral and Pailler, “one of the main constraints in building corpus from Twitter is the inability to obtain the entire stream, even for an extremely short period of time; there is not technical reason for this restriction, which actually results from the choices made when designing the site” [Dagiral and Pailler 2018, 117] (our translation). In order to have the entire stream, Twitter does offer the researchers to purchase data. However, this option strongly depends of available funding and raises some concerns about research ethics as well.

12 Next to the rather time-consuming tasks of building ones own dataset, researcher mostly rely on available data, which are shared online, often through platforms such as Github[5] or through special collections made available by archives, libraries, and other cultural heritage institutions. In 2010, the Library of Congress (LoC) reached an agreement with Twitter to archive all the tweets issued since 2006 [Raymond 2010]. It took the library until 2013 to complete the data collection for the period 2006-2010 – a fact demonstrating the enormous time and work required for indexing such a vast dataset. The volume of data (170 billion tweets and more than 300 terabytes of information from 2006 to 2010) led the LoC to announce in December 2017 that it would restrict collection to data concerning specific themes and events. For technical reasons but also ethical (consent) and legal (privacy rights) issues, data collected by the LoC are still not available to researchers [Zimmer 2015].

13 However, other institutions with smaller perimeters are offering access to Twitter collections. Examples include the archives of the INA (French National Institute for Audiovisual) and the BnF (French National Library), which are both related to the French Internet legal deposit. These institutions do not collect all the Twitter data, but have Twitter archives linked to their archiving perimeters. Although these archives are really valuable, they are biased on choices that do not necessarily coincide with the expectations of researchers. One respondent to our survey revealed the potential use of the French BnF archives, whilst underlining certain limitations: “I generally find it very difficult to work on traces, and particularly old ones. I use the BnF web archives, but the search tool isn’t very good. On top of that, anything that isn’t from ‘public’ pages isn’t archived” (ICS interview, May 2018). This may lead to research questions that are based on the availability of relevant archives (thus merely updating questions that researchers have already been aware of for some time). “The idea of the patrimonial value of web content” influences the criteria for safeguarding data and “threatens the existence of content that does not fit the criteria set by web archiving organisations for the preservation of data [...]. Certain contents and formats can also be excluded – this is particularly true of images in 1990s web archives, and also banner advertising and pop-ups” [Schafer 2018, 53].

14 This small overview of the ways researchers can access social networks sources (researcher's collect, use of sample made by others, purchase of data, institution’s collections) already underline several challenges in terms of sustainability and highlight several temporalities at stake:

- Researchers (and librarians) who collect data in real time in Twitter run the risk of lacking distance and potentially overlooking important trends. As explained by Thomas Drugeon, who works at INA in France and launched an emergency Twitter collection during the Charlie Hebdo attack in 2015:

  > I added hashtags to the collection at lunchtime on the day after the events in the editorial offices at Charlie Hebdo. We acted more quickly for the events of 13 November. The Twitter flows weren't organised in quite the same way either. The hashtags were more varied in November, whereas in January nearly everything was concentrated on the #jesuischarlie hashtag. In November, at least five hashtags emerged and we could observe movements, cycles too, for example day/night reflecting the different time zones at international level. For #jesuischarlie we missed the peak at the beginning. [...] Valérie Schafer: And did you also collect #jesuisahmed, #jenesuispascharlie? Thomas Drugeon: No, we didn't capture them ourselves, but we could access them indirectly, as a side effect. We didn't want to start adding hashtags along the way as we wanted to make sure that we would have a homogeneous set.[6]

- Researchers who react after an event will not be able to fully access data unless they buy them from Twitter or use datasets which were created by others (where available), may they be individuals or librarians. Researchers then face other challenges: although the datasets may be documented, stressing the importance of metadata [Pomerantz 2015], they are reliant on the choices of others – maybe even robots.
- As pointed out by Thomas Bottini and Virginie Julliard, who used INA archives to analyze the semiotic dimension of the "gender theory controversy" that played out on Twitter,[7] researchers need to "consider the often non-explicit archiving logic of these databases, as well as changes in their architecture and editorial policies" [Bottini and Julliard 2017, 47]. The massive availability of data in archiving collections is not synonymous with "turnkey" corpora [Barats 2016] [Brunet 2012] as collecting involves specific criteria and curation, which claim for a knowledge of archiving choices.

Coping with the sheer volume of data is a further challenge. The Info-RSN project, funded by the French National Research Agency (ANR), which focused on the circulation and sharing of information on Twitter and changes in journalism [Mercier and Pignard-Cheynel 2018], may help to illustrate this problem. The researchers used a private company to collect the data, i.e. all the tweets that contained a URL link to one of the 31 French news entities included in the survey. The raw data collected corresponded to roughly 40 million tweets, and it took one and a half years to develop tools and ultimately obtain a final corpus of 17 million tweets (after the elimination of duplicates and tweets that were outside the research criteria). 15

From a methodological point of view, building large datasets renews the traditional criteria mobilised for the constitution of research corpora (representativeness, exhaustiveness, homogeneity, etc.). The heterogeneous nature of data (alphanumeric, audio, visual or timestamped data), their ephemerality (including for the short term), the design-specific limitations of platforms (websites, socio-digital networks, internet portals), and finally the hardware-specific nature of devices and means of reception (computers, tablets, smartphones) make corpus-building a highly complex and critical hermeneutic practice. For researchers working with open access datasets, it can be difficult to detect the blind spots of a search query or to reconstruct the selection criteria that have informed the creation of the corpus. The archiving of Twitter by the BnF and INA can serve as a speaking example here again. INA compiles its archive through the public Twitter API, whereas the BnF uses Heritrix - the crawler also used by Internet Archive and many libraries around the world. The results are very different in terms of data as well as metadata: the BnF's Twitter archives look more like screenshots, whereas the INA collections are more flexible, allowing comparisons of metadata while giving access to results that no longer look like the original tweet. This may be a disadvantage or not, depending on the type of research being conducted: studying a Facebook or Twitter news feed from a semiotic and discursive point of view requires being able to retain the format of the broadcast. 16

Despite their apparent accessibility digitised texts also raise challenges. Some of them are common with born-digital heritage. Some databases are unstable over time. This is the case of private press databases like Factiva,[8] which 17 depend on agreements concluded between the publisher of the database and the owners of the data that is collected. Digitisation is not neutral either, and affects the digitised text [Föhr 2017] from a semiotic point of view:

> The source text [...] is then turned into various types of semiotic text, including a first, the “digital” text per se, in which it is encoded using 0s and 1s. This text is generally illegible for humans. A second, known as an “image text,” can be displayed on a screen or printed on paper and read as such, but any analysis of this version is typically “manual.” Then come the dynamic text, the annotated text and the edited text, and finally the readable, analysable text. So [...] digitisation does not produce a single “digital” copy of the text, but an entire galaxy of digitised texts which, interlinked and organised hierarchically, open up new avenues for interpretation and analysis. [Meunier 2018]

Factiva, for example, collects and digitises the texts without preserving the layout and format of distribution. These choices should be accessible to researchers whether they are stable or not. Etienne Brunet compared the Frantext[9] textual database with the French works on Google books. Frantext provides the researcher with a defined and explicit database for collection criteria [Brunet 2012].

Dealing with datasets as “sources” for humanities and social sciences research not only requires a critical understanding of the production and indexical logic of databases and its structured information, but also a hermeneutic reflection on the dual nature of digital data as both content *and* form. “When working with digital objects it’s essential to remember that what they look like on the screen is a performance,” says Trevor Owens [Owens 2020, 6], head of the Digital Content Management department at the Library of Congress in Washington D.C. Next to the challenge of determining the authenticity of data in terms of their physical integrity, one has to consider the performative volatility of data representations in different formats and on different devices. New skills such as interface criticism which make us think about the framing of “in-form-ation” [Drucker 2014] are therefore necessary in order to critically evaluate the “integrity of representation” of digital data [Föhr 2017]. 18

Finally, there is also the question of internal metadata and of metadata [Pomerantz 2015] added by the researcher during her/his annotations. Regarding the first ones, let’s mention the case of retweets. Although retweets are useful to study the popularity or even the virality of a tweet, the number of retweets that researchers may find in Web archives is linked to the exact time the tweet was archived, meaning that they may miss the real popularity of a given tweet if it was archived at the beginning of its dissemination and before the climax of its popularity. Regarding the second ones [Salvador 2017], the durability of taxonomies and their readability remain a central question: the choice of metadata and taxonomy is not always stable even within an organisation and this can make longitudinal or diachronic approaches more difficult, as the researcher has to consider not only changes in the Web, but also in its archiving, interfaces, supports, algorithms, descriptions and taxonomies. A rigorous documentation of metadata is therefore needed to facilitate searchability, analysis and quotations. 19

1.2 Long-term preservation, trial and error

Another question we addressed in our surveys is: what happens to the data at the end of the project? In the field of biology, Timothy H. Vines and his team conducted a survey on data sustainability through the analysis of 516 articles published from 1991 to 2011. They noted that scientific data are lost two years after the publication of an article, after which the chances of accessing scientific data would fall by 17% per year. This was mainly due to the failure of storage tools or obsolete email addresses that prevented any contact with the authors [Vines et al. 2014, 95]. Is the situation different in social sciences and humanities? Probably not! The answers to our survey show a great difficulty when it comes to implementing a sustainable storage and sharing approach in a medium and the long-term perspective (i.e. periods of up to ten years or beyond ten years, respectively). 20

A first finding of ours is, that testimonials show a focus on the “here and now” rather than on future uses of research data. The mastering of data and tools is often driven by a learning-by-doing approach, shifting the focus to pragmatic solutions rather than investing time into more fundamental speculations about open science in a sustainable 21 perspective. Many researchers highlight the large amount of short-term difficulties: “I don’t have much time to think about the long term, because I am too busy with my current collection and analysis problems! And above all, I don’t have the time or the technical means to go into it in detail” (ICS interview, May 2018). However, many of the scholars interviewed keep screenshots on their computers and external hard drives. They do their own archiving, but without sharing it or without any guarantee that they will be able to transfer their archives from one machine to another in order to preserve the data. This problem is sometimes “solved” by the use of online storage and sharing tools (OwnCloud, Dropbox, Google Drive, Huma-Num), “for practical reasons but without any real thought about the long term, especially after the project has ended” (ICS interview, July 2018).

But doing research is of course an iterative process that comes with a lot of surprises and unforeseen challenges which force the researcher to adapt methods and tools as the project progresses. It would be an error to underestimate the researcher’s adaptability and flexibility in this regard. As one historian confessed: “My current project (traces of the First World War on Twitter) was not originally designed as a research project. So initially, data archiving, storage and so on were not part of my plan. Today, they are increasingly central to my reasoning, as the project will end in 2019. For archiving, I can use the DH Now project. I also think that adapting to the GDPR is something to contemplate for the long term” (History interview, May 2018). 22

In terms of maintenance let’s return to the trio of data, corpus and results we underlined: they all can have similar stakes, but they involve different actors, research problems and management methods. Institutions such as the BnF preserve the data, without necessarily keeping the corpus created by the researchers. This complementary approach is currently being explored in the Internet Lab platform (made available to the researchers of the Web90 and ASAP projects in 2016), or in the prospective study conducted by Moiraghi (2018). Although data maintenance can be carried out by legal repositories and institutions, it can also be a private sector initiative, for example, for the Newsgroups of the Usenet community: these exchanges between early adopters of computer-mediated communications are a rich source of information for digital historians, and the maintenance of this data is entirely dependent upon Google’s interest in preserving these exchanges. Yet users were rapidly disappointed by the organisation of these archives, which were experienced as unclear and difficult to use and search [Poulsen 2009]. Camille Paloque-Bergès notes that “the non-standardised archives of the Web are the result of the succession of transitional and informal states of digitalised document cycles; this succession is due to the development of ‘wild’ documentation on the Internet before professional archivists started their work” [Paloque-Bergès 2017]. 23

The problem of citability of research data and final research outputs is of course not a new one. As Emmanuelle Bermès (BnF) noted, this concern existed way before the era of digitalisation: from 2006 onwards, the BnF was concerned about the sustainability of identifiers, notably for websites. A choice had to be made between “opaque identifiers” based on existing standards such as UUID (Universal Unique Identifier), or more “significant” identifiers that facilitated the citability and accessibility of resources. “This duality illustrates the problem libraries have always had: the dilemma of choosing between communication and conservation” [Bermès 2006] (our translation). There are further issues to address: “Indeed, sustainability and unicity both have a problem of scale. It is now commonly accepted that identifiers can be reused for different websites or even destroyed on the web. Sustainability cannot therefore be considered as ensuring that data is ‘eternal’, but rather that it ‘lasts long enough’ for the needs of the institution managing the resource” (idem). Sustainability therefore involves other challenges such as the independence of the authority maintaining the data or the ability to integrate pre-existing models and envision post-existing models. 24

Maintenance requires the consideration of both the temporality of the research questions, methods, and tools and of the stakeholders and users involved. It requires time, effort and resources that are rarely available to researchers if not planned from the outset. It is rare for the researcher’s institution to carry out this maintenance. Researchers may therefore decide to rely on *ad hoc* solutions. However, the question of the responsibility for data preservation remains: within an institutional framework, research is not the sole property of researchers. They could rightly argue that there is a mutual obligation of maintenance within the institution they have developed their research. Yet such solutions have been more often provided for article repositories and platforms than for research data. In this respect a system like HAL[10] in France offers an effective response that is increasingly used by researchers for making articles available through open access (preprints, author drafts, final publications, etc.). Similarly, sound archive repositories, like the 25

“Phonothèque” at the MMSH (the Maison méditerranéenne des sciences de l’homme, or Mediterranean Centre for Humanities),[11] offer a means of heritagising and preserving collections. But there are hardly any integrated solutions for projects using heterogeneous data. It is worth noting that various new solutions are under consideration at national level in France, for example the “feasibility study for a [shared] warehousing service for simple data.”[12] The way this study will deal with “Big Data, Little Data, No Data” [Borgman 2015] and with the precise scope of “simple data” remains to be seen, as in practice it is generally for complex data that researchers need help and support. But there are also question marks as to the need to create new solutions rather than building on existing ones proposed by research infrastructures in this area and to the utility of a nationwide scheme for an issue that is of international or global scope (not least because of cross-border research projects).

2. Conflicting temporalities and distributed materialities

As mentioned in the introduction, the different temporalities of research practices and preservation initiatives have to be problematised in order to tackle the problem of readability and intelligibility of research data in a long-term perspective. Dealing with digital sources or data in terms of sustainability requires taking into consideration the historicity of the “layered” or “distributed” materiality [Blanchette 2011] of digital datasets and objects, which is characterised by the interweaving of hardware and software environments. As Johanna Drucker has stressed, “the distributed concept requires attention to the many layers and relationships of hardware, software, bandwidth, processing, storage, memory, and other factors. The distributed approach registers a shift from materiality grounded in a single feature or factor to an approach based on multiple systems of interrelated activity” [Drucker 2013, Paragraph 21]. Although the interviews show that the colleagues interviewed take various dimensions of the distributed materiality of digital artefacts into account, the solutions are fragmented and often specific to individual researchers or a project, rather than shared and disseminated. 26

2.1 Tools’ evolution and obsolescence

The interviews we conducted indicate that researchers use a wide range of data analysis tools (Nvivo, Iramuteq, Gephi, etc.), and also reveal tools developed by researchers. Thomas Bottini and Virginie Julliard developed their own tool to observe the semiotic dimension of the controversy around gender studies on Twitter. This raises the question of the long-term registration of these “custom-made” tools, their diffusion and their readability in the medium term [Bottini and Julliard 2017]. This is true for the case of the Calico project (communities of e-learning, instrumentation and collaboration), which ran from 2006 to 2009 and examined distance learning platforms and online discussion lists [Blondel et al. 2011]. A set of analytical tools were developed during this project which are no longer available – be it because of professional mobility of researchers or the lack of funding to maintain the tools. This lack of sustainability directly affects the readability and the intelligibility of the results of this research project. 27

As noticed by J. Smithies and al. “digital projects benefit from being planned and executed with their longevity in mind from the start” [Smithies et al. 2019, 8]. The consideration of collection environments and conditions as well as the documentation work may enrich reflections on the choices of the observable elements [Fuchs and Angel 2018] and the choice of methodologies [Paveau 2014] [Venturini et al. 2014]. A good illustration of this point is an ANR-funded research project on online petitioning. The research group behind this project (Pluridisciplinary Analysis of On Line Petitioning) had access to data from one of the main French-speaking online petitioning websites (lapetition.be): the dataset included 12,000 petitions which had received a total of 3.25 million signatures over the period between 31 October 2006 and 12 February 2015. The richness of the data (data entered by Internet users and generated by the system), as well as their heterogeneity (alphanumeric data, number of signatories, site sections and forms, etc.), led to the development of different approaches, both qualitative (interviews with the designer, “manual” analyses) and quantitative (textual statistical tools), in order to look at the characteristics of online petitioning as well as the limitations and incentives induced by the system that affect the petitioning dynamic. The interviews proved to be essential in order to assess the stability of the system over time and avoid bias in interpreting the collected data. Before the end of the project, i.e. 2018, the website was no longer accessible, confirming the importance of taking screenshots in order to be able to contextualise the data [Barats 2016]. This example underlines the crucial importance of documentation – and the 28 emergence of data stewards as a new and promising profession in the field of digital humanities. As query tools evolve over time and can thus change data collection conditions, a professional documentation of the creation of research data should not only provide technical information (standards, protocols, formats), but must include search modalities used to constitute the corpora. The lack of “thick description” of both the metadata and the process of gathering, curating, and publishing or recontextualising research data is certainly one of the biggest risks when it comes to the long-term and sustainable contribution of digital humanities research to the long tradition of “analogue” knowledge production in humanities and social sciences [Tóth-Czifra 2020].

Until now, historians had been probably better protected from the obsolescence of their sources than their counterparts working with contemporary data, but they are now increasingly concerned by this risk. The ANR Web90[13] project carried out from 2014 to 2018 is a prime example: when it was launched in 2014, searches conducted in the Web archives (may it be the Wayback Machine of Internet Archive or collections of national institutions) only allowed URL requests. Subsequently, the BnF implemented a full-text search in its archives, thus widening the search capacities, affecting the results and influencing data analysis. In addition to changes in data access and the consequent disruption caused by full text and the possibility of classifying the results according to different facets (time, format, domain name), other additional tools provided an answer to a blatant problem: the temporally disparate composition of archived pages, which included anachronistic elements that had not been updated (for example, calendars that are not routinely re-archived for reasons of duplication, etc) [Brügger 2018]. Internet Archive recently introduced a system to date the archives of each element making up a webpage,[14] thus removing the opacity of these temporal patchworks. It is also important to document collection tools - they may sometimes exclude advertisings, gifs or emoticons in the case of automatic collection tools - or consultation terminals that modify the display of data, a central aspect in the case of works that emphasise a semiotic perspective. In the Web archives it should be noted that the fonts and characters may also differ from the original pages, as enlightened by Musiani et al.: “If at the time of archiving, a Web page font was not written explicitly in its original source code, but rather used by default, the default settings set by the browser in its current version will appear on the archived page” [Musiani et al. 2019]. Archiving data can also remove the original formatting of the platform. This is the case, for example, for Twitter data stored by INA as explained earlier: the form and the interface are not preserved, because the choices that have prevailed favour manageability and interoperability between the data. It is also important to note that the archiving or displaying of data can evolve over time: initially, INA archived emojis but did not display in the results to researchers, thus removing the heterogeneity and complexity of data that is characteristic of web writings [Halté 2016]. 29

2.2 Temporalities of users

One respondent noted that the main difficulty “is to reconstruct the temporal dynamics of the discussions studied (often I only have access to the final data)” (ICS interview, May 2018). Elements that have not been collected have to be reconstructed to account for the complexity of the *dispositif* [15] and its functionalities. This was undertaken by the INA, who collected data collection limit messages via Twitter’s public API and developed documentation on missing data. Research on the history of the Web has also shown that it is difficult (but not impossible) to identify the users of 1990s Newsgroups and establish their sociological profile, or even to consider the representativeness of their posts. The same question could arise for Twitter’s archives and its users. Preservation is also about the readability of the article in the future. Therefore, one should also consider future users. A publication may describe facts that may no longer be readable/understandable fifty years later, as the digital environment studied has evolved or even disappeared. If we imagine, for example, the disappearance of Twitter, how could we guarantee the documentation of the data in such a way as to reproduce their context and their techno-semiotic environment? This is already the case when reading an article about the Minitel - a videotext system broadly used in the 80s and 90s in France, with students, especially strangers to the French context, who have absolutely no knowledge of the technical ecosystem and the small beige box. Some current analyses can be difficult to understand even now: hyperlink analysis and maps can lack readability, for example in the field of scientific controversies when the data are not visible. This difficulty could be further accentuated if Gephi, Iramuteq or Lexico, are no longer usable, or even documentable. If the tools as well as the data disappear, is the analysis still readable? The design of the tools and *dispositifs*, studied for example by Monnoyer-Smith (2016) or Mabi (2016), introduce restrictions and develop a specific type of research and results (graphs, clouds, link maps 30

[Plantin 2013]), which also claim for digital hermeneutics.

In addition to the different temporalities of research data and the distributed materiality of the research devices and infrastructures, there are also temporal questions that are specific to the individual experiences of researchers. Amongst the researchers who were especially sensitive to data sustainability in our survey, one colleague – who has the ability to carry out “*ad hoc* programming and programming for the exploitation of XML databases” (DH and Language interview, May 2018) – has been particularly attentive to sustainability issues: 31

> In general, we deal with the volatility of the data on a day-to-day basis: we have transferred our servers three times. Part of the corpus data is hosted on personal servers, outside the research institutions. However, we are fortunate to have redundant servers in the laboratory that we manage internally: all our data are saved in XML format (txt) that is always accessible and guarantees their use in the long-term, and in a compressed form that goes from one server to another several times a day. In each of our projects, a substantial part of out funding is earmarked for 1 / the legal protection of the data, ensuring the sharing and the diffusion of the data entered (without confiscation) and 2 / the saving of data without a subscription (even if the data is hosted by a provider to avoid giving priority to one university project over another. No “cloud” storage is used, but hard drives are purchased and CRON backup programming is used on the websites of both project partners, etc.) Each time, the solution is to reinvent, according to the type of financing and the project itself: an editorial project? A “pure data” project? etc. (DH and Language Interview, May 2018)

The capacity to think about data maintenance and the constant reworking of ideas seems to develop as soon as the researcher is no longer in “discovery mode” and fully immersed into the difficult daily control of tools and data (as underlined by one respondent to our survey: “[...] It has been more difficult since I started trying to integrate source code into the analysis for critical code studies. However, I am interested in discourse, the practices and the uses - accessing the source code and becoming familiar with it can help to understand the active role code plays in human-machine interactions” (ICS interview, July 2018)). For those starting a DH-project with rather limited technical or data management skills, the need for support by computer scientists and engineers might be a necessity right from the start of a project. As underlined by a researcher in ICS: “Accessing (and processing) data seems increasingly dependent either on researchers learning IT methods that are not part of their initial background and that may be costly to acquire, or on working in conjunction with computer scientists (technicians, engineers or researchers) whose methods and research interests may be significantly different. We may therefore witness a narrowing of research aims or questions because of the need to master certain tools” (ICS Interview, May 2018). 32

3. Open Access and Digital Wastelands

National and international research infrastructures and institutions are seeking to provide answers to the transformations that challenge current research. The French Progedo[16] VRLI (Very Large Research Infrastructure for production and management of data) was created to preserve survey data. In the United States, Harvard University has initiated Dataverse,[17] a social science data storage project intended to facilitate Web access to verified web and non-web data and ensure their preservation [King 2007]. Similarly, to name but a few, the CESSDA (Consortium of European Social Science Data Archives) has added sustainability to its main objectives,[18] while OPERAS, a European Research Infrastructure for the development of open scholarly communication in the social sciences and humanities, and COPIM (Community-led Open Publication Infrastructures for Monographs) have recently been launched. 33

Despite such initiatives, at present, most born-digital heritage of HSS research is unequally preserved. While mailing lists and some research projects have been particularly well preserved,[19] a quick look at the archives of digital research projects on the Web reveals many broken links. Some researchers have tackled these issues, including Nanni (2017), who in “Reconstructing a website’s lost past - Methodological issues concerning the history of www.unibo.it” explains his efforts to rebuild the evolution of the website of the University of Bologna, which has lost much of its content. The challenges linked to the loss of devices or digital materials and software can no longer be ignored; this 34 includes CD Roms during the 80s and 90s, floppy disks, etc. Whilst emulation is often carried out for data archiving, it is rarely achieved for production and valorisation.

3.1 Sharing and Open Access

Open access is central to the concerns of researchers and of publishers, but is also of importance to European funders, especially in the framework of H2020 projects, thus continuing a movement that began with early initiatives such as ArXiv[20] in the field of physics or biology. Although this study does not intend to detail the policies of open access in Social and Human Sciences, it is important to note that these policies are a quintessentially economic and political challenges [Cotte 2017]. This is a regular subject of debate [Girard 2017] [Valluy 2017], which leaves researchers faced with a paradoxical demand: “first, to publish, then publish more, in highly rated journals, in journals that are increasingly evaluated, graded, ranked and second, to make results public, disseminate them to as many people as possible, get them online as quickly as possible” [Farchy et al. 2010] (our translation). 35

Open access has found its way through platforms like the French multidisciplinary open archive HAL[21] or the research notebooks platform Hypotheses.[22] These collaborative efforts have definitely fostered best practices and paved the way for a greater acceptance of open access publication and data sharing. Sure, one could argue that the maintenance and the searchability of these platforms aren’t free from biases and institutional dependences too. However, these biases and dependences appears less worrying than the ones specific for private companies with their opaque and questionable system of ranking, visibility and referencing. Beyond the independence of these academic databases, attention needs to be paid to questions of interoperability of data formats. But like the open access, interoperability comes at a price regarding policies of standardisation, negotiations and debates they may raise, technical developments and implementation, updates, gateways, etc., and the human effort it implies. The question of the economic model is clearly posed when funded researchers can publish articles in open access by paying publishers, while others are prevented from doing so because they are unable to pay the high fees.[23] 36

The question of open access is crucial at all stages of research: it entails strong constraints (as well as opportunities) for the researcher before, during and after the research process, and requires not only specific means (research engineers, data collection tools and short- and long-term storage) but also a detailed knowledge of the rules applying to data sharing (which were for example renewed within European legislation by the recent GDPR - General Data Protection Regulation).[24] These rules also evolve over time, when some works enter the public domain. Controversy may sometimes appear, as seen in the debates between [Ertzscheid 2015] and the Anne Frank Foundation. The researcher who currently uses web archives from archival institutions may for her/his part be frustrated by the legal impossibility to copy archives due to intellectual property rights. But he/she can find comfort in the knowledge that they will be traceable through permalinks. 37

As Zimmer (2010) pointed out in the case of Facebook, this sharing and quotability raise issues of anonymity. Anonymity can be waived by search engines and betrayed by the cited data – despite the best intentions of researchers [Latzko-Toth and Pastinelli 2013]. It is therefore necessary to anticipate any obstacles to anonymisation, which vary according to the type of data. This difficulty was encountered by the group of researchers working on the ANR-funded project on online petitioning [Barats 2016].[25] As mentioned earlier, thanks an agreement with the manager of an e-petitioning platform, this project had access to the entire database and related data (including texts, petitioners, signatories, comments and timestamps) over a period from 2006 to 2016. The data had been previously anonymised (12,000 petitions and more than 3 million signatories). However, the researchers quickly realised that entering the quotation of comments by signatories via the Google search engine provided the same information online. The research team had to carry out multiple online checks to be certain before publication that no quotes or excerpts compromised the anonymity of the data, thus restricting the dissemination of results. In addition, the agreement did not allow them to respond to requests for data sharing. This example is typical of the limits of open access to digital data and of the ethical, technical and legal constraints that influence the dissemination of results. 38

3.2 “404 not found” and other challenges raised by time

39 Unfortunately, many great digital humanities projects of the past now suffer from their status as “digital wasteland.” Many research project websites are just abandoned at the end of the project; others may be frozen. However, freezing a project is not sufficient to maintain it or guarantee its sustainability. A good example is the Amsterdam-project DDS (*De Digital Stad*): an online socializing platform created in Amsterdam in the mid-1990s. A team of media archeologists around Gerard Alberts revived the DDS platform using the data that had been preserved, but identified a number of technical issues despite the precautions previously taken:

> On 15–16 January 1996, the DDS servers were down for most of the night to allow for a full backup of De Digitale Stad. A full 1-on-1 disk copy of all the servers running DDS was created on 3 Digital Linear Tapes (DLT). DDS congratulated itself with a city frozen in time, preserved “to be studied by archaeologists in a distant future: the FREEZE.” [...] In restoring old data, it soon came to light that the package would not simply unwrap, or defrost. The DLT tapes holding the FREEZE did not easily render their content. [Alberts et al. 2017]

The planning of the DDS founders here was remarkable. However, being prepared is not always sufficient to avoid risk. One survey respondent described how the earmarking of funds to store data beyond the end of an international research project was still insufficient to keep the data available on the medium term: “The data was hosted by a university institution which could no longer cover the costs of data storage, leading to the loss of all the data. We had planned the financing of data maintenance, but the funding was limited to a ten-year period and could only be temporarily extended, leading to the loss of data that could have been used for future research” (ICS interview, June 2018).

40 As we know, many projects stop abruptly or are no longer updated. Although their short-term fate could be considered better than that of the many websites that simply disappear, the high risk of the data disappearing has resulted in multiple preservation efforts. For example, the European Publications Office keeps information about European projects within the CORDIS database,[26] and the websites of these projects can provide extremely useful data for funding bodies, stakeholders or historians interested in the history of science or of the European Union. This determination is also shown by the arquivo.pt website of the Portuguese web archives, which has made an effort to preserve European R & D websites.[27]

41 Some past projects give the feeling that research productions are out of date, but they reflect the state of the digital art of the period: a pioneering 3D production in 1980 rapidly appears to be well below current levels of achievements. Something that was a technical feat in its time becomes one of the first steps of a technique (which may be extremely relevant for analysis by a historian of science). This brings us to a wider examination of obsolescence over time for certain projects and the evaluation of the maintenance efforts necessary to avoid it: this raised the issue of digital databases or research outcomes based on paid images and royalties that must be negotiated on a regular basis.

42 It also highlights the broader notion of updatism [Pereira and Lopes de Araujo 2016], which seems inherent to the conflicting temporalities of digital studies. Yet the updating of websites and data are different things. The second case can be illustrated by the Twitter archives concerning the terrorist attacks that occurred in France over the last few years. These archives continue to grow, particularly at INA, as the hashtags are reused by users for tweets about different terrorist attacks or for commemorations, thus continuously producing new results. While living archives are more and more developing [Rhodes 2013], updatism is a new challenge that may affect the perimeters, the temporalities of projects and even reviews (possibly claiming for post publication reviews[28] as a way to reflect the evolving nature of data and knowledge).

Conclusion

43 “Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs.” This definition of the concept was formulated by the Brundtland Commission in 1987, as commissioned by the UN.ℍ.[29]

This definition of sustainable development could be a useful line of thinking in the field of human and social sciences, even if it was not designed for this domain. Let us take the current example of INA's policy to homogenise the data and metadata resulting from the archiving of socio-digital networks (for example between Twitter and Facebook). As explained by Jérôme Thièvre (INA Web Legal Deposit), web archivists aim to create common metadata to access Twitter and Facebook archives, despite being faced with two different environments and *dispositifs*. But their work has two different goals: beyond the homogenisation of this data / metadata to meet the current needs of consultation and cross-checking, they also consider keeping other metadata in the "back office," behind the scenes, that are not currently available to the researcher but which could be made accessible and used for future needs [Schafer 2018]. 44

By making the researcher's "toolbox" readable and explaining her/his "ways of doing things" [Lécossais and Quemener 2018, 8], the working environment and the temporality of the research are documented for the long term. This sometimes requires a certain curiosity for aspects that can be understated, obscured or considered of little value in the production of knowledge but could yet play a role in reinforcing the reflective approach that is central to research. The question of the long duration of research in comparison to the short cycles of digital innovations and the rapid evolution of devices, tools and *dispositifs* requires to include thinking about sustainability into our critical tradition of hermeneutics. Many examples in our article are based on born-digital data, although some issues are common to retro-digitised archives or digitised collections. However, all these cases involve a number of different temporalities (that of data, infrastructure, tools, devices, hardware and software, media, research, institutions and funding) and multi-stakeholder issues that require collective reflection to clearly identify the actors and locations that are best adapted to implement and support the challenges of data sustainability. These challenges are deeply transdisciplinary, as evidenced by the responses of researchers in ICS, history, sociology, and linguistics. 45

Yet they are also interdisciplinary: they require an answer that calls for archivists and librarians, ethical and legal approaches and involvement of disciplinary expertise. They also claim for awareness-raising and training for researchers, who describe the constant tension of dealing with current challenges whilst anticipating future issues in research, infrastructure, ethical, legal and political terms. This necessary interdisciplinarity poses many difficulties, such as the development of a common language that goes beyond technical jargon, as illustrated by the difficulties the Bamboo project faced in the United States [Dombrowski 2014, 3]. Finally, these challenges call for taking into account the highly internationalised research framework and several environments, such as national data protection regulations, where access and protection need to be reconciled, and where regulations can also make certain agencies obsolete. All these questions are strongly linked to constantly changing legal matters, which are sometimes interpreted in a way, as seen in the narrow opposition between the right to memory and the right to be forgotten that Dulong de Rosnay and Gadamuz (2017) deconstructed. 46

This is also a social and political challenge that brings to light not only funding, but also capitalisation on existing structures and scientific legacy to avoid the duplication of means, supports and tools. From here onwards, what was seen as the constraint of sustainability becomes the challenge to attain efficiency, sharing and responsibility,[30] aiming to also address the inherently ethical issues that could be framed as "sustainability ethics" [Kibert et al. 2011].[31] Although Smithies and al. noticed that "maintenance and archiving of digital scholarship is an iterative, continuous process, that does not allow for perfect endings" [Smithies et al. 2019, 13], this maintenance, which is well studied in infrastructures such as transport [Denis and Pontille 2015], should be better considered and studied in the domain of knowledge infrastructures. We fully agree with Russell and Vinsel (2018) and the maintainers' group, who speak out against the fascination with prioritising innovation over maintenance. Instead of promoting the rhetorics of newness and digital revolution, one should pay more attention to the danger of digital obsolescence and the massive fading away of digital data and knowledge. 47

Notes

[1] https://digitalmatters.utah.edu/digital-matters-lab/

[2] See https://www.huma-num.fr/presentation/reseau; https://www.dariah.eu/; https://operas.hypotheses.org.

[3] https://library.stanford.edu/research/data-management-services/data-management-plans

[4] 10 colleagues are historians, 5 are researchers in Information and Communication Studies, 3 in linguistics and 2 in sociology. Despite disciplines, all colleagues were concerned with data access and data preservation for individual analysis as well as collective projects. Quotations were translated from French to English.

[5] See for example https://ruebot.net/post/twitterdatasets/. Nick Ruest shares Twitter IDs, not tweets.

[6] https://asap.hypotheses.org/173#more-173. Interview conducted during the Asap project dedicated to the born-digital heritage of Paris terrorist attacks, in 2016 (our translation).

[7] As underlined by their abstract, “this paper sets out the method developed in a study that explores the production of Gender during the controversy over ‘Gender theory,’ on Twitter. The method articulates semiotics and computer science to further our understanding of how the techno-editorial properties of this micro-blogging platform influence that production.” The controversy emerged after the adoption in May 2013 of a law allowing same-sex marriage. It has sparked much debates, some of them linked to gender studies, whose opponents depicted as a “gender theory.”

[8] https://professional.dowjones.com/factiva/

[9] This database includes more than 5000 references to French works or texts from the 10thto the 21st century, and over 251 million words. See https://www.frantext.fr/.

[10] https://hal.archives-ouvertes.fr/

[11] http://phonotheque.mmsh.huma-num.fr/dyn/portal/index.seam;jsessionid=ba99be050dd4cea60bb98fbe5e32?page=home&

[12] https://www.ouvrirlascience.fr/principes-et-lignes-directrices-de-locde-pour-lacces-aux-donnees-de-la-recherche-financee-sur-fonds-publics/

[13] https://web90.hypot heses.org/le-projet-web90

[14] See the article posted by the Internet Archive foundation (05/10/2017): “Wayback Machine Playback… now with Timestamps!” https://blog.archive.org/2017/10/05/wayback-machine-playback-now-with-timestamps/

[15] The French term *dispositif*, as defined by Michel Foucault, is sometimes translated as apparatus. See Callewaert (2017): “I use the French term *dispositif*, even in English, because I feel the usual translations such as ‘apparatus’ or ‘deployment’ are misleading when used in connection with Foucault’s texts. These terms contained different meaning in the trend of philosophy in France at the time of the development of the term dispositif.”

[16] http://www.progedo.fr/progedo/

[17] https://dataverse.org/about

[18] See https://www.cessda.eu/About

[19] See for example http://liste.theuth.fr/ (archives of this mailinglist are available from 2004). See also the astonishing stability and durability of a founder public history project: The Valley of the Shadow, launched from 1993. http://valley.lib.virginia.edu

[20] https://arxiv.org

[21] https://hal.archives-ouvertes.fr/

[22] https://fr.hypotheses.org/

[23] See https://blogs.mediapart.fr/edition/au-coeur-de-la-recherche/article/130418/transition-vers-l-acces-libre-le-piege-des-accords-globaux-avec-les-editeur. Interesting resources on this broad debate are also available at: http://blogs.lse.ac.uk/impactofsocialsciences/.

[24] https://ec.europa.eu/info/law/law-topic/data-protection/reform_fr

[25] http://textopol.u-pec.fr/anr-appel/ and http://textopol.u-pec.fr/anr-appel/index.php/2017/09/19/podcasts-de-la-journee-journee-appel-sur-le-

cadre-juridique-applicable-aux-traitements-de-donnees-a-caractere-personnel/

[26] https://cordis.europa.eu

[27] See “Arquivo.pt preserved websites about Research & Development projects funded by the EU” on aquivo.pt (10/04/2017). http://sobre.arquivo.pt/en/arquivo-pt-preserved-websites-about-research-development-projects-funded-by-the-eu/

[28] Appraisal and revision of a paper occurs or may continue after publication.

[29] https://sustainabledevelopment.un.org/content/documents/5987our-common-future.pdf

[30] See the COMETS document (CNRS)https://www.univ-paris13.fr/wp-content/uploads/guide_promouvoir_une_recherche_inte_gre_et_responsable_8septembre2014.pdf.

[31] On sustainability ethics: “The ethics of sustainability provides a clear sense of the principles that make sustainability more than just a simple problem-solving system, but make it an idea that is grounded in commonly understood ethical principles. In short, the ethics of sustainability provide the moral authority behind sustainability as a fair and equitable approach to making the world a better place.” http://rio20.net/wp-content/uploads/2012/01/Ethics-of-Sustainability-Textbook.pdf.

Works Cited

Alberts et al. 2017 Alberts, G., Went, M. and Jansma, R. “Archaeology of the Amsterdam digital city; why digital data are dynamic and should be treated accordingly”, *Internet Histories*, 1, 1-2 (2017): 146-159.

Bachimont 2017 Bachimont, B. *Patrimoine et numérique. Technique et politique de la mémoire*. INA Editions, Bry-sur-Marne (2017).

Badouard 2016 Badouard R. “‘Je ne suis pas Charlie.’ Pluralité des prises de parole sur le web et les réseaux sociaux”. In P. Lefébure and C. Sécail C. (eds) *Le défi Charlie. Les médias à l’épreuve des attentats,* Lemieux Editeur (2016), pp. 187-219.

Barats 2016 Barats, C. (ed) *Manuel d’analyse du web*. Armand Colin, Paris (2016).

Barats et al. 2016 Barats, C., Dister, A., Gambette, P., Leblanc, J-M., Peres-Leblanc, M. “Analyser des pétitions en ligne: potentialités et limites d'un dispositif d'étude pluridisciplinaire”, *JADT* 2016, *Actes des Journées internationales d'Analyse statistique des Données Textuelles* (2016). http://jadt2016.sciencesconf.org

Bermès 2006 Bermès, E., “Des identifiants pérennes pour les ressources numériques. L’expérience de la BnF”, *BnF* (2006) http://www.bnf.fr/documents/ark_presentation_bermes_2006.pdf.

Blanchette 2011 Blanchette, J-F. “A material history of bits”, *Journal of the American Society for Information Science and Technology* (2011) https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.215420.

Blondel et al. 2011 Blondel, F-M., Giguet. E. “Analyses et partages de corpus de discussions avec calico- leçons tirées d’une expérience récente”. *EPAL - Echanger Pour Apprendre en Ligne*, Jun Grenoble, (2011) https://hal.archives-ouvertes.fr/hal-02010626/document.

Borghoff et al. 2006 Borghoff, U., Rödig, P., Scheffczyk, J., Schmitz L. *Long-Term Preservation of Digital Documents, Principles and Practices*. Springer, London (2006).

Borgman 2015 Borgman C. L. *Big Data, Little Data, No Data*. MIT Press, Cambridge MA (2015).

Bottini and Julliard 2017 Bottini T.,Julliard, V. “Entre informatique et sémiotique. Les conditions techno-méthodologiques d’une analyse de controverse sur Twitter”, *Réseaux*, 2017/4: 35-69.

Brunet 2012 Brunet, E. *Au fond du GOOFRE, un gisement de 44 milliards de mots*. Conferences Lexicometrica (2012) http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Conferenciers-invites/Brunet,%20Etienne%20-%20Au%20fond%20du%20GOOFRE.pdf.

Brügger 2018 Brügger, N. *The Archived Web. Doing History in the Digital Age*. MIT Press, Cambridde MA (2018).

Callewaert 2017 Callewaert, S. “Foucault’s Concept of Dispositif”, *Praktiskegrunde*, 1-2 (2017): 29-52.

Clavert 2018 Clavert, F. “Commémorations, scandale et circulation de l’information: le Centenaire de la bataille de Verdun sur Twitter”, *French Journal for Media Research*, 10 (2018) http://frenchjournalformediaresearch.com/lodel/index.php?

id=1620.

Cotte 2017 Cotte, D. “Économies scripturaires, formes documentaires et autorité. Réflexions et esquisse d’analyse des architextes de la ‘science ouverte’”, *Communication & langages*, 2017/2: 117-129.

Dacos 2012 Dacos, M. “Manifeste des Digital Humanities” (2012) https://tcp.hypotheses.org/318.

Dagiral and Pailler 2018 Dagiral, E., Pailler, F. “Des chercheur·e·s et des tweets. Enquêter sous contraintes”. In S. Lécossais, N. Quemener, (ed), *En quête d’archives: bricolages méthodologiques en terrains médiatique*, INA Éditions (2018): pp. 113-21.

Denis and Pontille 2015 Denis, J., Pontille D. “Beyond breakdown: two horizons of maintenance work”, *i3 Working Papers Series*, 15- SES-08 (2015).

Dombrowski 2014 Dombrowski, Q. “What Ever Happened to Project Bamboo?”, *Literary and Linguistic Computing*, Volume 29, Issue 3, (2014): 326-39 https://doi.org/10.1093/llc/fqu026.

Drucker 2013 Drucker, J. “Performative Materiality and Theoretical Approaches to Interface”, *Digital Humanities Quarterly* 7/1 (2013) http://digitalhumanities.org/dhq/vol/7/1/000143/000143.html.

Drucker 2014 Drucker, J. *Graphesis. Visual Forms of Knowledge Production*. Harvard University Press, Cambridge (2014).

Dulong de Rosnay and Guadamuz 2017 Dulong de Rosnay, M., Guadamuz, A. “Memory Hole or Right to Delist?”, *RESET*, 6 (2017) http://journals.openedition.org/reset/807.

Ertzscheid 2015 Ertzscheid, O. “Anne Frank et le domaine public. Mon combat?”, *Affordance.info* (2015) http://affordance.typepad.com//mon_weblog/2015/10/anne-frank-mon-combat-.html.

Farchy et al. 2010 Farchy, J., Froissart, P. and Méadel C. “Introduction – Sciences.com, libre accès et science ouverte”, *Hermès* 2, n° 57 (2010): 9-12.

Fickers 2017 Fickers, A. “Digital History : On the heuristic potential of thinkering”. Keynote lecture at DH-Nord Lille (2017) https://publi.meshs.fr/ressources/digital_history_on_the_heuristic_potential_of_thinkering.

Fickers 2020 Fickers A. “Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur digitalen Forensik?”, *Zeithistorische Forschungen/Studies in Contemporary History* (2020): 157-68 https://zeithistorische-forschungen.de/1-2020/5823.

Fuchs and Angel 2018 Fuchs, C., Angel, C. (eds) *Organization, Representation and Description through the Digital Age, Information in Libraries, Archives and Museums*. De Gruyter, Berlin (2018).

Föhr 2017 Föhr P. “Historische Quellenkritik im Digitalen Zeitalter”, Doctoral Thesis, University of Basel, Faculty of Humanities and Social Sciences (2017).

Girard 2017 Girard, C. “Les mécanismes de centralisation des données de la recherche”, *Revue française des sciences de l’information et de la communication*, n°11 (2017). http://journals.openedition.org/rfsic/3255.

Halté 2016 Halté, P. “Enjeux pragmatiques et sémiotiques de l’étude des émoticônes”, *Réseaux*, vol. 197-198, no. 3 (2016): 227-52.

Hoekstra and Koolen 2018 Hoekstra, R., Koolen, M. “Data Scopes: Towards Transparent Data Research In Digital Humanities”, DH2018, Mexico (2018) https://dh2018.adho.org/data-scopes-towards-transparent-data-research-in-digital-humanities/.

Kibert et al. 2011 Kibert, C. J., Thiele, L., Peterson, A., Monroe, M. “The Ethics of Sustainability”, *Free Textbook List* (2011) http://www.freetextbooklist.com/the-ethics-of-sustainability/.

King 2007 King, G. “An Introduction to the Dataverse Network as an Infrastructure for Data Sharing”, *Sociological Methods and Research*, 36 (2007): 173-99.

Lang et al. 2012 Lang, D. J., Wiek, A., Bergmann, M., Stauffacher, M., Martens, P., Moll, P., Swilling, M., Thomas, C. J. “Transdisciplinary research in sustainability science: practice, principles, and challenges”, *Sustainability Science*, 7, (2012): 25-43.

Latzko-Toth and Pastinelli 2013 Latzko-Toth, G., Pastinelli M., “Par-delà la dichotomie public/privé: la mise en visibilité des pratiques numériques et ses enjeux éthiques”, *tic&société*, Vol. 7, 2 (2013) http://journals.openedition.org/ticetsociete/1591.

Lécossais and Quemener 2018 Lécossais, S., Quemener, N. (ed). *En quête d'archives: bricolages méthodologiques en terrains médiatique*. INA Éditions, Bry-sur-Marne (2018).

Mabi 2016 Mabi, C. "Analyser les dispositifs participatifs par leur design". In C.Barats (ed.) *Manuel d'analyse du web en Sciences Humaines et Sociales*, Armand Colin (2016), pp. 33-37.

Mercier and Pignard-Cheynel 2018 Mercier, A., Pignard-Cheynel, N. (eds) *#info Commenter et partager l'actualité sur Twitter et Facebook*. Éditions de la Maison des sciences de l'homme, Paris (2018).

Merzeau 1998 Merzeau, L. "Ceci ne tuera pas cela", *Les cahiers de médiologie*, vol. 6, 2 (1998): 27-39.

Meunier 2018 Meunier, J. G. "Digital text: hermeneutic issues", *Digital Humanities Quarterly*, vol. 12, n°1 (2018) http://www.digitalhumanities.org/dhq/vol/12/1/000362/000362.html.

Moiraghi 2018 Moiraghi, E. "Étude prospective sur les besoins et les attentes des futurs usagers du Laboratoire d'étude et d'analyse de corpus numériques", Paris, *BnF* (2018) http://actions-recherche.bnf.fr/BnF/anirw3.nsf/e7cfeb857fb88bebc1257b52004c05f9/cb0177ed072cd5fdc1258264004fdf0b/$FILE/DSR-CORPUS_Etude-de-besoins_Janvier2018_RGB.pdf.

Monnoyer-Smith 2016 Monnoyer-Smith, L. "Chapitre 1 - Le web comme dispositif: comment appréhender le complexe?". In C. Barats (ed.), *Manuel d'analyse du web en Sciences Humaines et Sociales*, Armand Colin (2016), pp. 11-31.

Mons 2018 Mons, B., *Data Stewardship for Open Science. Implementing FAIR Principles*. CRC Press, Boca Rota (2018).

Musiani and Schafer 2017 Musiani, F. and Schafer V. "Patrimoine et patrimonialisation numériques", *RESET*, 6 (2017) http://journals.openedition.org/reset/803.

Musiani et al. 2019 Musiani, F., Paloque-Bergès, C., Schafer, V. and Thierry, B. "Qu'est qu'une archive du Web? OpenEdition", *Marseille* (2019) https://books.openedition.org/oep/8737?lang=en.

Nanni 2017 Nanni, F. "Reconstructing a website's lost past – Methodological issues concerning the history of www.unibo.it", *Digital Humanities Quarterly*, vol. 11, n°2 (2017) http://digitalhumanities.org:8081/dhq/vol/11/2/000292/000292.html.

Owens 2020 Owens T., "Digital Sources and Digital Archives: The Evidentiary Basis of Digital History". In D. Staley (ed), *A Companion to Digital History*, (2020 forthcoming) https://doi.org/10.31235/osf.io/t5rdy.

O' Sullivan 2019 O' Sullivan, J. "The humanities have a 'reproducibility' problem". *Talking humanities* (2019) https://talkinghumanities.blogs.sas.ac.uk/?s=O%27+Sullivan&submit=Search.

Paloque-Bergès 2017 Paloque-Bergès, C. "Vers des lieux de mémoire réticulaires?", *RESET* 6 (2017) http://journals.openedition.org/reset/839.

Paveau 2014 Paveau, M-A. "L'alternative quantitatif/qualitatif à l'épreuve des univers discursifs numériques", *Corela*, HS-15 (2014) http://journals.openedition.org/corela/3598.

Pereira and Lopes de Araujo 2016 Pereira, M., Lopes de Araujo V. "Historical time reconfigurations: presentism, updatism, and loneliness in digital modernity", rev. Ufmg (Revista da Universidade Federal de Minas Gerais), *belo horizonte*, v.23, n.1-2 (2016): 290-97 https://www.ufmg.br/revistaufmg/pdf/REVISTA_23_web.pdf.

Plantin 2013 Plantin, J.C. "Chapitre 11 - D'une carte à l'autre: Le potentiel heuristique de la comparaison entre graphe du web et carte géographique". In C. Barats (ed.), *Manuel d'analyse du web en Sciences Humaines et Sociales*, Armand Colin (2013): pp. 228-245.

Pomerantz 2015 Pomerantz, J. *Metada,* MIT Press, Cambride MA (2015).

Poulsen 2009 Poulsen, K. "Google's abandonned library of 700 million titles (updated)", Wired (2009) https://www.wired.com/2009/10/usenet/.

Raymond 2010 Raymond, M. "How Tweet It Is!: Library Acquires Entire Twitter Archive", *Blog of the LoC*, (2010) https://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/.

Rhodes 2013 Rhodes, T. "A Living, Breathing Revolution : How Librarians Can Use 'Living Archives' to Support, Engage, and Document Social Movements", IFLA Wlick 2013, Singapore (2013) https://thescholarship.ecu.edu/bitstream/handle/10342/4496/084-rhodes-en-TamaraRhodes.pdf?sequence=1&isAllowed=y.

Russell and Vinsel 2018 Russell, A. Vinsel, L. "After Innovation, Turn to Maintenance", *Technology and Culture*, 59, 1

(2018), pp. 1-25 https://muse.jhu.edu/article/692165/pdf.

Rygiel 2011 Rygiel, P. “L’enquête historique à l’ère numérique”, *Revue d’histoire moderne & contemporaine*, vol. 58-4bis, no. 5 (2011): 30-40.

Salvador 2017 Salvador, “Indexer des documents « du dedans » : quels moyens de répondre à la question du lieu de la donnée (XML, OWL, RDF,REST)?” In J. Longhi (ed.), *Communication*, 35/1, (2017).

Schafer 2018 Schafer, V. “Les réseaux sociaux numériques d’avant...”, *Le Temps des médias*, 2, 31 (2018): 121-36 https://www.cairn.info/revue-le-temps-des-medias-2018-2-page-121.htm.

Smithies et al. 2019 Smithies, J., Westling, C., Sichani, A-M., Mellen, P. and Ciula, A. “Managing 100 digital humanities projects: digital scholarship & archiving in King’s Digital Lab”, *Digital Humanities Quarterly*, 13, 1 (2019) http://www.digitalhumanities.org/dhq/vol/13/1/000411/000411.html.

Tóth-Czifra 2020 Tóth-Czifra, E. “The Risk of Losing the Thick Description. Data Management Challenges Faced by the Arts and Humanities in the Evolving FAIR Data Ecosystem”. In J. Edmonds (ed), *Digital Technology and the Practices of Humanities Research*, Open Book Publishers (2020), pp. 235-66.

Valluy 2017 Valluy, J. “Libre accès aux savoirs et accès ouvert aux publications”, *Revue française des sciences de l’information et de la communication*, n°11 (2017) http://journals.openedition.org/rfsic/3194.

Venturini et al. 2014 Venturini T., Cardon D. and Cointet J.P. “Présentation, Méthodes digitales Approches quali/quanti des données numériques”, *Réseaux*, vol. 188, 6 (2014): 9-21.

Vines et al. 2014 Vines, T. H., Albert, A.Y.K., Andrew, R. L., Débarre, F., Bock, D., G., Franklin, M. T., Kimberly J. G., Moore, J-S., Renaut, S. and Rennison, D.J. “The Availability of Research Data Declines Rapidly with Article Age”, *Current Biology*, vol. 24, 1 (2014): 94-7.

Wilkinson et al. 2016 Wilkinson, M., Dumontier, M., Aalbersberg, I., et al. “The FAIR Guiding Principles for scientificdata management and stewardship”, *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18.

Zimmer 2010 Zimmer, M. “‘But the data is already public’: On the ethics of research in Facebook”, *Ethics and Information Technology*, 12, 4 (2010): 313-25.

Zimmer 2015 [Zimmer 2015] Zimmer, M. “The Twitter Archive at the Library of Congress: Challenges for Information Practice and Information Policy”, *First Monday*, Vol. 20, 7 (2015) http://firstmonday.org/ojs/index.php/fm/article/view/5619/4653.