Towards a sustainable social media archiving strategy for Belgium

---

WP2 Report

Analysis of user requirements (Task 2.2)

Editors	Peter Mechant & Eveline Vlassenroot
Responsible partners	MICT (UGent) KBR
Version	1.0
How to cite this?	*P. Mechant & E. Vlassenroot, BESOCIAL: Analysis of user requirements (Task 2.2), July 2021.*

Table of Contents

1. Introduction

This report describes the results of the BESOCIAL Task 2.2, titled 'Analysis of user requirements'. The goal of this task is to gain insights in the needs and requirements of a broad range of stakeholders (researchers, cultural heritage professionals, publishers, policy-makers and other potential end users) when using a social media archive. These insights will be taken into account when designing (access to) the social media archive itself. This task also wants to understand how specific user requirements of potential social media archive users might influence adoption and usage of such archives. Meeting these requirements and sociotechnical needs is not an easy task; already in 2000 a persistent 'sociotechnical gap' between users' or researchers' requirements and the functionalities or affordances of available interfaces and (technological) infrastructures has been identified (Ackerman, 2000).

Research for this task encompassed three phases. After an initial desk research phase (i), we decided to use qualitative semi-structured interviews (ii) with potential end users of social media archives to gain insights in the needs and requirements of a broad range of stakeholders. Insights and take-aways from these semi-structured interviews were then validated in a pre-conference workshop (iii) of the 2021 RESAW conference. Below, we discuss the results of these three phases.

2. Desk research

Some user requirements studies on the needs of potential users of (social media) web archives have been conducted in the past. Important take-aways were amongst others:

- that a significant segment of the research community in the humanities and social sciences is still unaware of web archives and that many do not know exactly what they contain or how they can be used (Costea, 2018).
- that there is a lack of awareness surrounding the existence of (inter)national web archives Riley and Crookston (2015).
- that potential scholarly users expressed wishes for visualisation tools, insight into the establishment of selection criteria and themes for special collections, the opportunity to nominate special collection themes and the inclusion of more images, blogs and rich media in the collection.
- that more light should be shed on the decisions surrounding selection policies and criteria by means of documentation, that tools for searching and analysing should be improved and that more communication about and promotion of the web archive is necessary (BnF, 2011).
- that most users do not restrict searches by date, that URL searches are common but that full-text search is preferred, that a lot of users search for names of people, places or things and that image search was identified as an information need that was not yet met (Costa and Silva, 2010).
- that providing more metadata to researchers, further developing user-friendly and shareable tools that are modular so that they can be easily adapted to different research disciplines, providing clear information about copyright issues and building communities that include both web archivists and researchers, is necessary (Dougherty et al. 2010).

More recent studies looked at how actually working with currently available web archives can provide more insight into what is needed for ensuring that (social media) web archives get adopted by academia (e.g. Jackson, Lin, Milligan, & Ruest, 2016; Ogden & Maemura, 2021; Ruest, Lin, Milligan, & Fritz, 2020).

For example, Ogden and Maemura (2021) compared and contrasted their experiences of undertaking web archival research at the UK Web Archive[1] and Netarkivet[2]. They invoked three conceptual devices (orientating, auditing and constructing) to describe common research practices and associated challenges and highlighted the significant time and energy required on the part of researchers to begin using national web archives, as well as the value of engaging with the curatorial infrastructure that enables web archiving in practice. Rather than presenting a linear workflow or fixed set of practices, their concepts *orientating, auditing* and *constructing* necessarily overlap, thus reflecting the complex, iterative and often exploratory processes involved in the development of web archival research projects.

- *Orientating* to the web archive includes engaging with web archives as new ontological devices for historical research; unpicking the often complex legal constraints of access; and embracing new ways of knowing data and infrastructure.
- *Auditing* the web archive includes engaging with the particularities of the collection and search interfaces of web archives; contextualising data by tracing a history of collection practices and curation decisions; and probing the limits and edges between data, collections and infrastructure.
- *Constructing* encompasses activities surrounding the creation of a subset of data to work with through more focused analyses. This includes negotiating and navigating the technical infrastructure to access diverse and varied forms of data; selecting and aggregating data from sources across 'collections'; and iteratively revisiting the possibilities of particular research methods given data availability.

Also, Ruest et al. (2020) point to the process of scholarly inquiry as the most critical component in building and sustaining a community around web archiving. They suggest a process model that decomposes scholarly inquiries into four main activities: *filter, extract, aggregate,* and *visualize*, stating that each of the activities need not occur in the same session or even the same location, and perhaps most importantly, with the same tools:

- *Filter* - Focusing on a particular subset of the web archive; this can be accomplished by content, metadata, or some extracted information.
- *Extract* - After selecting a subset of material, the scholar typically then extracts some information of interest. Examples include extracting the plain text from the raw HTML source, identifying mentions of named entities or assessing the sentiment of the underlying text.

[1] https://www.webarchive.org.uk

[2] https://www.kb.dk/en/find-materials/collections/netarkivet

- *Aggregate* - The output (a collection of records of interest) needs to be aggregated or summarized before they are suitable for human consumption.
- *Visualize* - The aggregated results are presented in some sort of visualization for the scholar's consumption.

3. Semi-structured interviews

3.1 Methodology

We decided to use qualitative semi-structured interviews to poll the needs and requirements of a broad range of stakeholders (researchers, cultural heritage professionals, publishers, policy-makers and other potential end users) when using a social media archives.
Semi-structured interviews allow for more flexibility in which topic lists do not need to be followed religiously and can be modified depending on the expertise or the issues raised during the conversation. They enabled us to approach the problem from within the context of the research subject and reveal both factual knowledge as well as the opinions of the interviewees.

Of course, the data collection is dependent on the will of the interviewee and the circumstances in which the interview takes place (such as location, time limitations etc.) can have an influence. However, this method is most appropriate to complement desk research and provide more insight into some sensitivities that may be at play and that would otherwise have gone undiscovered. Also, organizing the interviews around a number of topics allowed the respondents to talk freely and at length and created the necessary space to allow them to give meaning to their experiences in working (or in envisioning working) with archived social media content.

In order to recruit interviewees for the semi-structured interviews described above, we used persona outlines. A persona is a fictional character created to represent a user type that might use a social media archive in a similar way. A persona is thus a composite representation of prevalent qualities of a user segment and will not exactly match a specific person or comprehensively describe the full diversity of a group.

In the literature we found some persona-descriptions that could function as inspiration and a guideline. For our personas we were inspired by:

- The personas created in the Corpus project and outlined in the deliverable 'Le projet Corpus et ses publics potentiels. : Une étude prospective sur les besoins et les attentes des futurs usagers'[3]. In this deliverable, five personas are described in order to help with

[3] https://hal-bnf.archives-ouvertes.fr/hal-01739730/document

"(…) l'identification et à la définition des profils des usagers potentiels" (p. 38). Similar to our effort, the Corpus-personas were 'mapped' on a digital literacy scale (Figure 1).

- The personas created to establish a more robust and data informed understanding of the individuals that engage digitally with the US National Archives[4]. This list includes personas such as the Genealogist, the Curious Explorer, the Educator, the Records Manager, the History Enthusiast, the Museum Visitor, the Researcher and the Veteran.

Figure 1: overview personas Corpus-project (https://hal-bnf.archives-ouvertes.fr/hal-01739730/document)

Based on the above, and taking group discussions and the results of the PROMISE project into account, five persona were created that could be used to start the recruitment process;

- Bart - post-doc researcher in communication sciences
- Febe - PhD-student in computational linguistics
- Ben - 35-year old journalist working for the newspaper
- Manou - 28-year old scientific researcher @ KBR
- Jan - 68-year old former team coordinator

A full description of these 5 personas is provided in Annex. Using these five persona as an 'ideal' we explored our personal and professional networks to find suitable interviewees and contacted them using e-mail. In total 6 interviews were conducted:

- 15.03.2021: interview with young digital humanities scholar exploring opportunities of social media archives (best aligns with persona 'Manou')
- 17.03.2021: interview with PhD computational linguistics (best aligns with persona 'Febe')

[4] https://www.archives.gov/digitalstrategy/personas

- 25.03.2021: interview with PhD candidate in communication sciences (best aligns with persona 'Bart')
- 12.04.2021: interview with an experienced digital humanities scholar (best aligns with persona 'Manou')
- 26.04.2021: interview with a retired manager of a business unit at an international company (best aligns with persona 'Jan')
- 30.04.2021: interview with a professor in literature (best aligns with persona 'Bart')

3.2 Interview guide

All participants were interviewed using conference call software (due to the covid19 restrictions). The interviews were semi-structured, using both pre-defined and open questions. Each interview was tape-recoded with permission, transcribed and analysed. The semi-structured interview guide encompassed 3 main themes:

- Selection: questions on what kind of content interviewees would like to see harvested and archived, to what extent, with which frequency etc.
- Consultation: questions on how interviewees would like to explore, consult and analyse content in a social media archive
- Involvement & support: questions on how interviewees envision their involvement in social media archiving strategies and how they would like to be supported by the archiving institute

3.3 Results

3.3.1 Selection policies of content to be archived

We received a variety of answers when we polled the interviewees about what content they think a social media archive should store and make accessible. This is not surprising as most interviewees, quite logically, referred to their own expertise and knowledge domain. For example, the computational linguistic we talked to showed interest in having social media archives available that could provide him with text data from certain specific interest or knowledge domains (e.g. the specific word use or language spoken by car enthusiasts and car dealers) in order to train domain specific language models, while the digital humanities scholar

would want to use archived social media content for the qualitative content analysis of the communication of specific social media accounts.

Similar to the interviewees of Stirling et al. (2012) the six respondents we spoke to realized that it is 'impossible to predict what material will interest professional or amateur researchers in the future' and they agreed that some degree of selection should be legitimate given the volume of data that exists. Some of them also suggested some ways to cope with this paradox and suggested (a stronger) inclusion of the academic field in selection decisions and policies (see further);

"Deciding what to archive seems very important to me. In order to engage academic researchers with collections, a first step should be to involve academics in these selection decisions or at least make them aware of which selection criteria are used.".

3.3.2 Exploring and selecting content from the archive

Despite diverse interests in *what* should be archived, most interviewees agree on *how* the archived content should be query-able. They state that an archive should enable a flexible selection approach by offering a diverse set of *selection affordances*, including but not limited to keyword, date range, social media platform, language, full text or e.g. hashtags search. In this context, one interviewee stated:

"Shouldn't or couldn't we expect a similar service such as the one offered by Google?".

In contrast two interviewees (with significant more knowledge and expertise in archiving then the others) remarked that a normal search interface or organization of search results (as an ordered list of hits) is probably inadequate or even impossible. This aligns with the remark by Costa and Silva (2010) who stated that: "In web search engines, the users' intents are mainly informational, then transactional and lastly, navigational. In web archives, the users' intents are mainly navigational, then informational and lastly, transactional." or with Jackson et al.'s (2016) suggestion to create an exploratory interface based on visualizations (according to Shneiderman's mantra for visual information seeking).

Another respondent, for whom one of the main selection criteria would be to have content created between certain dates remarks that, if searching by date is limited to the date the site was crawled, this selection criteria is not useful for her as she needs filtering or search results within a particular publication date.

When asked to explore the archived TweetSets collection of the George Washington University[5] and remark on the *selection affordances* offered through its interface, all respondents expressed that they experienced the interface as very transparent and offering a wide range of parameters to select (create) a dataset. Also the provenance information ('source dataset'), the descriptive statistics ('dataset statistics) and the examples ('Sample tweets') that are provided with each dataset were experienced as very useful. Similarly, the option to generate a preview, based on one's selection criteria ('Get a sample of tweets in the dataset and see dataset statistics') was described by many as a 'good example of best practice'.

One specific interviewee remarks in this context that:

"Providing good and exhaustive descriptions of the available metadata and giving sufficient information about the dataset should be common practice. Moreover, this is also important in terms of democratizing data; not everyone has the necessary skills or competences, so I like the fact that this website [the archived TweetSets collection of the George Washington University] provides generic descriptive statistics on the dataset. Now, everyone can at least grasp what the data is about.".

Some respondents also expressed the need to gain insight in *what is actually missing* from and in certain collections and suggest to add this type of metadata to archival collection descriptions:

"For me it would be beneficial to also have information on what was not captured, to which depth a page's hyperlinks were followed (or not).".

Indeed, access constraints, as well as current search interfaces make it difficult to see the collections 'in the round' or from a vantage point that gives a sense of where the boundaries of the archive lie. Here, tools such as Periphery could play a role allowing collection owners to define how missing resources are expressed during the replay of archived content (Brucker, 2020).

In general interviewees stated that 'one should aim to provide as much information on any given collection as possible'; interestingly for some this should also include references to certain methodologies or certain software or toolsets that can be used to analyse the collection as well as to articles or research papers that (partially) used the collection in the past:

"I am a complete newbie in the field of social media archiving so the biggest hurdle that I face is rather basic and fundamental, that is: to grasp what is possible, to understand the opportunities that these archived collections make possible… In that context it would be really

[5] https://tweetsets.library.gwu.edu/datasets

useful for me to have examples or pointers available towards research that has been done in the past on these collections".

3.3.3 Involvement and support

While some interviewees did not know in detail the current practices used to collect social media content, most of them indicated that they would be interested to be involved in creating the selection policies for this content. Interviewees proposed various approaches such as collecting and archiving social media accounts that are most visited or most linked, social media accounts that are deemed important at a given time, or social media accounts that are collaboratively decided upon by domain-specific working groups.

Most respondents agree that the potential role of a Belgian social media archiving community as well as cooperation around this topic between all stakeholders is key. As such they resonate Stirling et al. (2012) whom stressed the need to engage with researchers and amateurs working with the web and to involve them in the creation of methodology and identification of sources within the archives. Here KBR Lab[6] could play an important role in involving the research community. KBR Lab could encourage researchers to make use of the resources contained in web archives, e.g. by communicating clearly and consistently when new archives are made available, by supporting researchers to the fullest extent possible or by organizing workshops or 'how-to's' focused on specific collections or on specific methodologies and tools. As such, KBR Lab could take the lead towards developing collaborative environments for researchers working with the sociotechnical infrastructure of Belgian (social media) web archives.

Although most respondents recognised the resource constraints under which most web archiving institutes operate, the value of curators engaging with researchers directly at the earliest stages of research was something that most of our interviewees often stressed during the interviews. However, respondents also clearly indicated that a web archive should first and foremost focus on its' *core* tasks while striking a balance with the needs of potential users of the web archive (especially those with less digital or computational skills).

3.3.4 Exporting and analyzing content

Most interviewees were not familiar with web archiving and had a rather naive idea on how they would be able to download relevant archived content and use it for their analysis. For example, when we pointed out that the result set of a selection from the archived TweetSets collection of the George Washington University boils down to a list of Tweet-identifiers (due to legal reasons

[6] https://www.kbr.be/en/projects/digital-research-lab/

no other information can be provided) or that for consulting some of these social media archives researchers need to be physically present at the archiving institution, reactions were often surprised or disappointed.

When discussing results that could be exported, respondents often neglected (or were unaware) of the big data features of most social media archives. Afterall, big data is "high–volume, –velocity and –variety information assets that demand cost–effective, innovative forms of information processing for enhanced insight and decision making" (Sicular, 2013). Also, legal constraints and the output format, were often not considered by the interviewees. Some of them for example indicated that they would not be able to technically 'rehydrate' the result set of a selection from the archived TweetSets collection of the George Washington University, or they stated that next to the textual content they would also expect the look-and-feel of social media channels to be archived.

"Ah ok, … if I understand you correctly it is not that simple. In that case I would definitely need some technical support because I am not familiar with using API's or dedicated 'big data' software."

In this context, we also noticed that the desired output formats vary; while some would need a structured-data export (e.g. json or csv) – with various degrees of available metadata – others expressed the need for actual screenshots or pdf-files that also capture the look-and-feel of the social media content. Several respondents pointed out that this is also linked to the toolset of each specific discipline and that researchers have specific needs that often cannot be met with general tools (or data formats).

4. Personas and RESAW Pre-conference workshop

4.1 Methodology

On 14/06 we organized a pre-conference workshop for the 2021 RESAW conference[7] titled 'User requirements for working with social media web archives' together with researchers from WARCNet (Web ARChive studies network)[8]. Organisers were Peter Mechant & Eveline Vlassenroot (imec-MICT-UGent), Sally Chambers (Ghent Centre for Digital Humanities), Niels Brügger (Aarhus University), Susan Aasman (University of Groningen), Friedel Geeraert & Fien Messens (Royal Library of Belgium, KBR).

The workshop was structured in four parts. First Sally Chambers and professor dr. Niels Brugger quickly introduced and describe BESOCIAL and WARCNet to the participants in order to introduce the context. Next, we had three interesting speakers lined up who will offered the participants some insights on what it means to try to preserve social media content and on what it takes to use social media content as research data. The last part constituted of an hands-on workshop that took about an hour and that used online break out rooms and an interactive visual dashboard on Miro. For this last part we had about 15 international participants with profiles ranging from archivist, preservation officer or information manager to post-docs and scientific researchers.

For the hands-on workshop participants were ask to pick a BESOCIAL persona and to take on the role of one of the given BESOCIAL personas and to reflect, discuss and elaborate on the hurdles, but also opportunities, that this persona would encounter when accessing or working with archived social media data. These 5 personas were then introduced to the participants (they were also provided with a full one-pager on the persona that they selected) and they were also briefly instructed on how to use Miro[9] in case they were not familiar with this online visual collaboration platform.

Once the break-out sessions of about 7 persons each were started, each group was also asked to pick a (social media) web archive as a case to discuss. They could opt for e.g. the institution one of the participants was working for, a best-practice example, or one of the archives provided by us, namely:

[7] https://www.resaw2021.net/
[8] https://cc.au.dk/warcnet/
[9] https://miro.com/app/dashboard/

- https://tweetsets.library.gwu.edu/datasets
- http://webadmin.oszk.hu/solrwayback/
- http://www.nationalarchives.gov.uk/webarchive/
- https://www.webarchive.lu/covid-19/
- http://data.webarchive.org.uk/opendata/ukwa.ds.2/#issues
- https://netpreserve.org/projects/collaborative-collections/

In the end, one of the participating groups decided to focus on the use case of http://data.webarchive.org.uk/opendata/ukwa.ds.2/ while the other group opted for the use case of https://tweetsets.library.gwu.edu/datasets. Finally, participants were asked to discuss and reflect on this use-case (from the viewpoint of their persona) for three separate phases (orientating / auditing / constructing). These three conceptual phases or devices describe common research practices and associated challenges when undertaking web archival research (Ogden and Maemura, 2021). A two-folded approach was used; first, to write down hurdles or challenges for that phase & discuss these, next, to answer the 'How might we solve this?'-question for the identified hurdles or challenges (individually and in group). The break-out sessions reconvened after about 45 minutes to plenary share their insights.

4.2 Results

4.2.1 Orientating

First we asked the participants to reflect on the *orientating* phase. Ogden and Maemura (2021) created this conceptual device to describe common research practices and associated challenges when engaging with web archives as new ontological devices for historical research. It involves getting to know the archiving (e.g. unpicking the often complex legal constraints of access) and embracing new ways of knowing data and infrastructure.

Results show that most personas encounter various hurdles from the beginning and that only the persona of Febe has actually all the necessary skills to cope with the selected web archives as new ontological devices. She finds https://tweetsets.library.gwu.edu/datasets easy to navigate and thinks that the condition of harvesting are clearly exposed and finds the descriptive statistics to get an insight on the collection, helpful. Other personas such as Bart or Jan are discouraged by the size of the collections (containing millions of tweets) or disappointed because only English language tweets seem to be provided and only the tweet-id can be downloaded. Orientating for the second group (who had chosen for the use case of http://data.webarchive.org.uk/opendata/ukwa.ds.2) proved to be harder. Here, even Febe, who

needs social media content specifically, found it difficult to identify/isolate this in the UKWA dataset. Also, Febe needs to test code for sentiment/subjectivity analysis, but it is not clear to her how to extract text for computational analysis in the UKWA dataset. Other hurdles such as the difficulty in understanding the legal deposit restrictions (persona Ben) or the provided meta data categories (persona Bart) were also noted.

Potential solutions to mitigate these problems in the *orientating* phase included; creating a universal search interface over all collections (although this already exists – at least for Twitter – see https://catalog.docnow.io/); providing suggestions for tools to use (e.g. for rehydrating tweet-ids); producing small (video) tutorials or harmonizing the interface design and offering the same "tools" for each archive. For Ben, an ideal situation would include "A kind of shop window with added links to a search interface, or alternatively a contact point to get in touch with an archivist e.g. via a chat box (during opening times)."

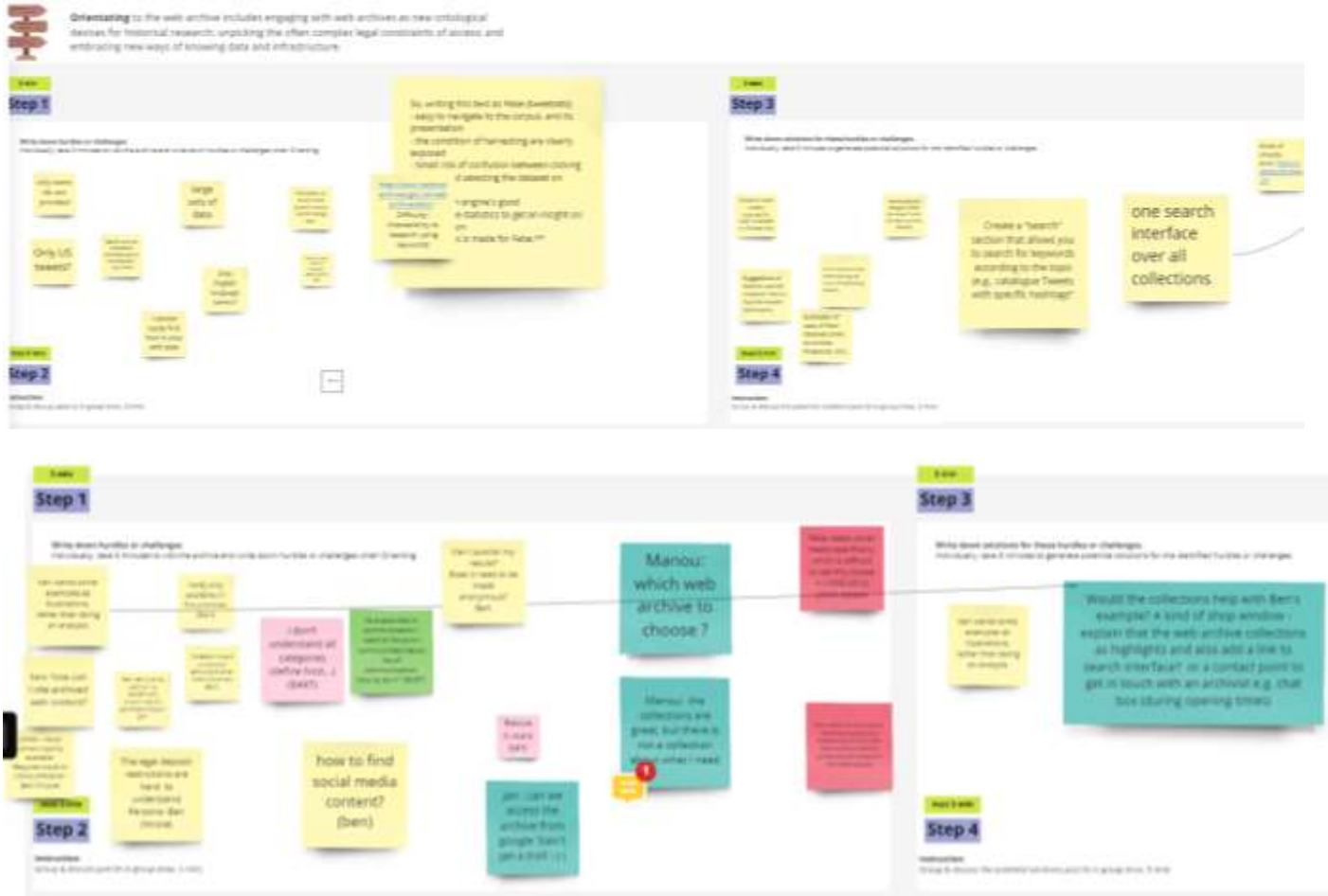

Figure 2: Overview of some of the workshop results on 'orientating'

4.2.2 Auditing

Next we asked the participants to discuss hurdles and potential solutions in the *auditing* phase (Ogden and Maemura, 2021). Auditing refers amongst others to engaging with the particularities of the collection and search interfaces of web archives. It involves contextualising data by tracing a history of collection practices and curation decisions. It requires probing the limits and edges between data, collections and infrastructure.

Results on the *auditing* phase show that most personas encounter mainly practical or semantical problems. For example, they had issues with locating the single-point-of-contact details for a given collection or experienced divergences between concepts and researched words. Persona Bart for example wondered how exactly he could extract tweets from the archive.

Potential solutions that were mentioned here include better displaying of information and better explanation of the articulation between tweet-sets and the 'dataverse'; creating a common thesaurus (e.g. by applying Natural Language Processing); or creating other links to web archives or a link to the Memento service.

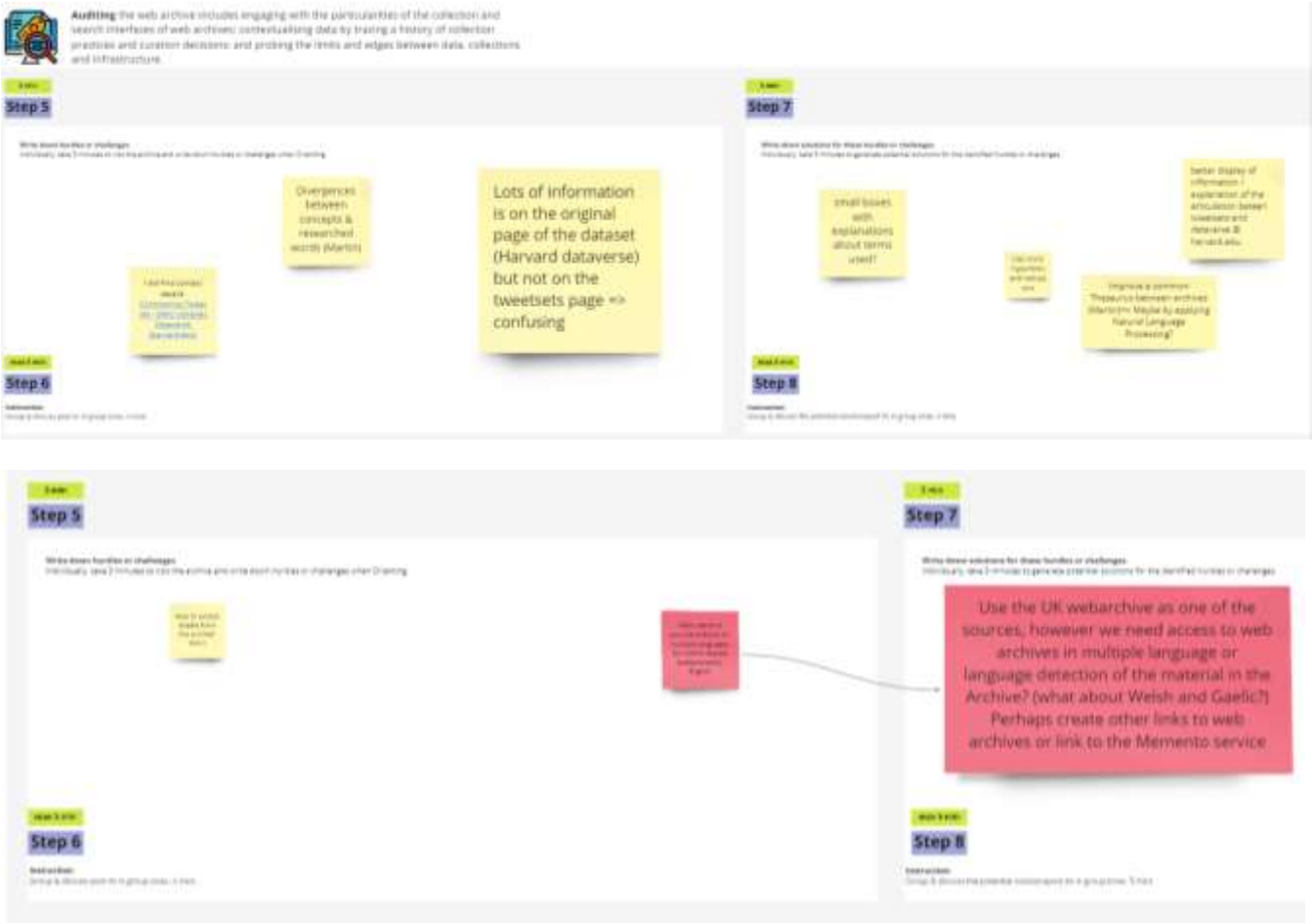

Figure 3: Overview of some of the workshop results on 'auditing'

4.2.2 Constructing

Lastly, we asked the participants to reflect on the *constructing* phase. This conceptual device can be used to describe activities surrounding the creation of a subset of data to work with through more focused analyses(Ogden and Maemura, 2021). This includes negotiating and navigating the technical infrastructure to access diverse and varied forms of data; selecting and aggregating data from sources across 'collections'; and iteratively revisiting the possibilities of particular research methods given data availability.

Due to lively discussions in both groups on the *orientating* and *auditing* phases both groups unfortunately lacked the time to discuss the *constructing* phase in an in-depth manner. Nevertheless, various challenges were remarked by the participants ranging from a lack of information on the collection (e.g. information on what was not collected), to the need of searching Chinese ideograms (in the case of persona Jan), to the problem that all datasets look separated for the case of https://tweetsets.library.gwu.edu/datasets, even when the archive probably has tweets in common; this requires then more work for potential users to download several datasets, merge them, remove duplicates, etc.

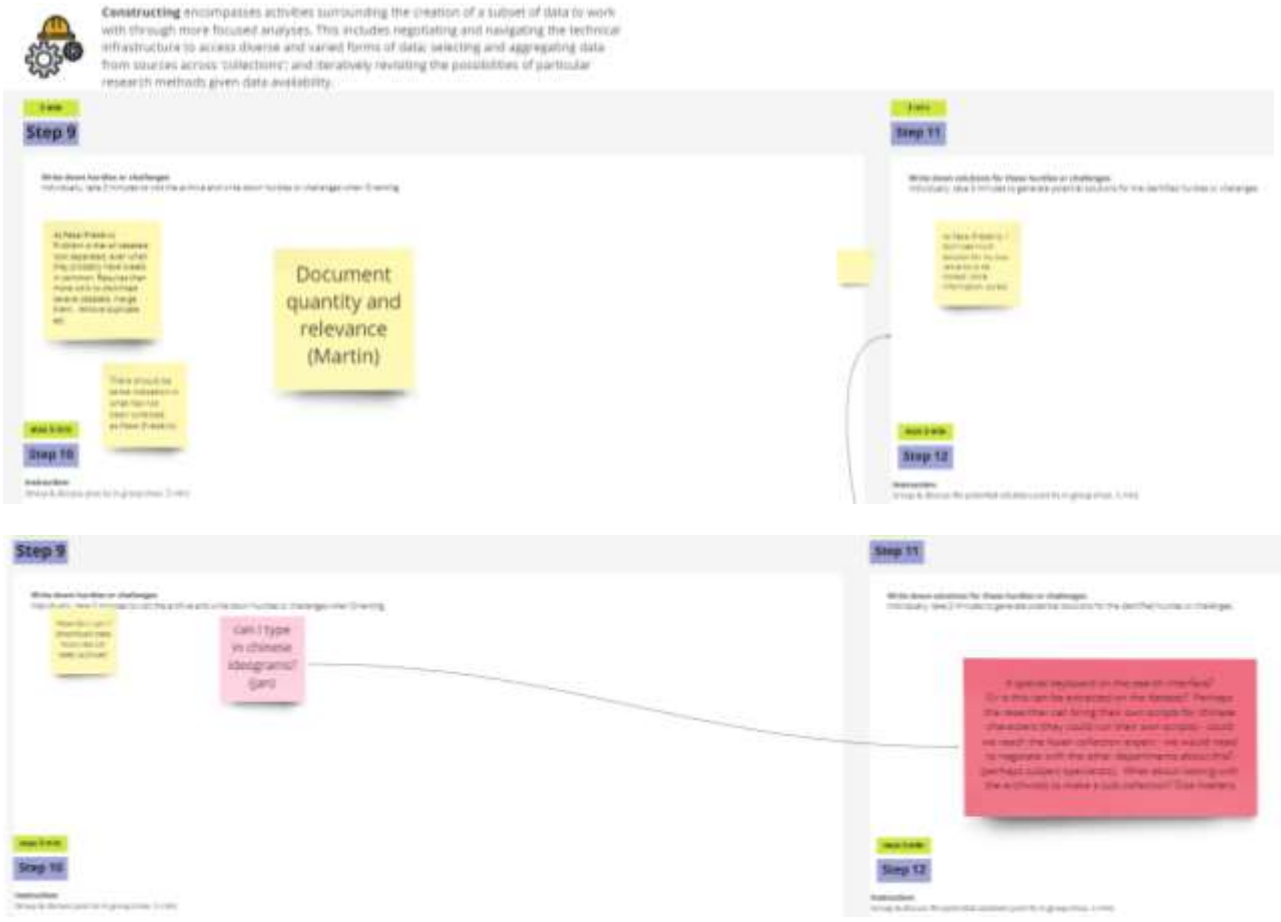

Figure 4: Overview of some of the workshop results on 'constructing'

5. Discussion and conclusion

In a recently published article Ogden and Maemura (2021) state that web archives are in many ways antithetical to research because they are composed of born-digital, previously online materials that are rendered largely inaccessible online (in archival format) due to the very mechanisms that enable their collection (e.g. legal deposit).

Although the results described in this report are based on a small number of interviews and one workshop we believe that they provide some interesting insights that can partially remediate this antithetical nature of web archives to research as argued by Ogden and Maemura (2021).

However, most notably, there still is a **lack of awareness surrounding the existence of (inter)national (social media) web archives**. More communication about and promotion of the web archive is necessary. We think that it is important that archiving institutes should realise that there still is a lack of awareness surrounding the existence of (inter)national web archives. The fact that a significant segment of the research community in the humanities and social sciences is still unaware of web archives is also something that became apparent in other research projects by Riley and Crookston (2015) or Costea (2018) and that could be mitigated by heightening the awareness of potential users via more targeted communication and marketing activities. In short, more communication about and promotion of the web archive is necessary.

Also, often **researchers are not aware of what archives exactly contain or how they can be used**. A solution here could be to supplement the meta- and provenance data of archived collections with references or links to papers or articles that have used the collection as research data, or to link up to suitable methods and software that could be used to analyse a collection. Our interviewees not only requested clear and exhaustive descriptions for each archived collection (outlining provenance info and such) but stated that they are also interested in a description of what the specific collection does not contain (for example listing the websites or social media channels that could not be crawled.)

In terms of selecting which social media content to archive (and which not) we argue for (a stronger) **inclusion of the academic field in selection decisions and policies**. As such we resonate Stirling et al. (2012) whom stressed the need to engage with researchers and amateurs working with the web and to involve them in the creation of methodology and identification of sources within the archives. Here, KBR Lab could play an important role in involving the research community. KBR Lab could encourage researchers to make use of the resources contained in web archives, e.g. by communicating clearly and consistently when new archives are made available, by supporting researchers to the fullest extent possible or by organizing workshops or 'how-to's'

focused on specific collections or on specific methodologies and tools. As such, KBR Lab could take the lead towards developing collaborative environments for researchers working with the sociotechnical infrastructure of Belgian (social media) web archives.

Despite the fact that diverse interests in what should be archived exist, we note that most interviewees **agree on how the archived content should be query-able**. However, the idea of classic search interface would often not suffice and further development of interfaces that open up and make searchable archived social media should be explored (e.g. an interface based on visualizations). This aligns with the remark by Costa and Silva (2010) who stated that: "In web search engines, the users' intents are mainly informational, then transactional and lastly, navigational. In web archives, the users' intents are mainly navigational, then informational and lastly, transactional." or with Jackson et al.'s (2016) suggestion to create an exploratory interface based on visualizations according to Shneiderman's mantra for visual information seeking (see e.g. Ahlberg & Shneiderman, 2003).

In general interviewees stated that 'one should aim to provide as much information on any given collection as possible'. Interestingly for some this should also include references to certain **methodologies or certain software or toolsets** that can be used to analyse the collection as well as to articles or research papers that (partially) used the collection in the past. KBR Lab could play an important role in involving the research community by encouraging researchers to make use of the resources contained in web archives and by supporting researchers to the fullest extent possible. This is important as researchers often lack the necessary (digital) skills or domain knowledge or neglect (or are unaware) of the big data and legal features of most social media archives.

6. References

Ackerman, M. S. (2000). The intellectual challenge of CSCW: The gap between social requirements and technical feasibility. *Human–Computer Interaction*, *15*(2–3), 179–203.

Ahlberg, C., & Shneiderman, B. (2003). Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In *The craft of information visualization* (pp. 7–13). Elsevier.

Brucker, M. (2020). Expressing Boundaries of Web Collections. Retrieved from Conifer project website: https://blog.conifer.rhizome.org/2020/08/10/periphery.html

Costa, M., & Silva, M. J. (2010). Understanding the information needs of web archive users. *Proc. of the 10th International Web Archiving Workshop*, *9*(16), 6. Citeseer.

Costea, M.-D. (2018). *Report on the Scholarly Use of Web Archives*. NetLab.

Jackson, A., Lin, J., Milligan, I., & Ruest, N. (2016). Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities. *IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, 103–106. https://doi.org/10.1145/2910896.2910912

Ogden, J., & Maemura, E. (2021). 'Go fish': Conceptualising the challenges of engaging national web archives for digital research. *International Journal of Digital Humanities*, 1–21.

Riley, H., & Crookston, M. (2015). *Awareness and use of the New Zealand web archive: a survey of New Zealand academics*. National Library of New Zealand and Victoria University of Wellington.

Ruest, N., Lin, J., Milligan, I., & Fritz, S. (2020). The archives unleashed project: technology, process, and community to improve scholarly access to web archives. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 157–166.

Sicular, S. (2013). Gartner's Big Data definition consists of three parts, not to be confused with three 'V's. *Forbes*. Retrieved from http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs

Stirling, P., Chevallier, P., & Illien, G. (2012). Web archives for researchers: Representations, expectations and potential uses. *D-Lib Magazine*, *18*(3/4). Retrieved from http://www.dlib.org/dlib/march12/stirling/03stirling.html

7. Annex

7.1 Persona #1: Bart

- **Age & Profession**: 42-year old post-doc researcher in communication sciences

- **Quote**: "The CfP ends in a few days, nevertheless let's try to submit an abstract. We will figure out how to actually gather the data and do the research when we get accepted."

- **Short description**: Bart lives with his wife and two teenage children not far from his work desk at the university. He likes meeting up with friends, reading science-fiction novels, binging YouTube content and the occasional run in order to stay fit.

- **Goal for using social media archive**: Bart has convinced his younger colleague who is working on his PhD to submit an abstract for a special call for paper. The abstract got accepted and now they need to urgently start researching the hypothesis that 'more extreme political parties post more 'populist' content/messages on Twitter'. In order to do that they need the tweets of all Belgian political parties tweeted over the last year.

- **Professional situation**: Bart has been working for more than 15 years at a research group that conducts user research in the domain of ICT & innovation and that is affiliated to the university's department of communication sciences. Although he has some programming expertise, acquired more than a decade ago, he feels that he is not up to date. Moreover, because of a heavy workload, he feels that he lacks the time to thoroughly learn/acquire new skills that are required. Similarly, although he has worked with big datasets in the past, he has little to no expertise in working with big data nor expertise in working with analytical software that can process those big datasets.

7.2 Persona #2: Febe

- **Age & Profession**: 24-year old PhD-student in computational linguistics

- **Quote**: “The party is not over until it is over”

- **Short description**: Febe lives with her girlfriend in a small studio apartment not far from work. She likes to party and to travel. She has been playing the saxophone since she was a kid. Most of her free time goes to performing music with several bands of which she is a member.

- **Goal for using social media archive**: Febe is involved in a European research project (that partially finances her PhD) in which she is trying to develop a methodology for sentiment analysis and subjectivity detection through the deep semantic analysis of text published via social media channels. In order to train and test her code, she is looking for a big data set with (textual) social media content. Given the European context of the project, Febe needs to analyse and develop the methodology for several languages.

- **Professional situation**: Last year, Febe started working as a PhD-student at a research group that focuses on fundamental and applied research in the domain of language and translation technology. In her first year, colleagues with more expertise helped her a lot with mastering the programming techniques and analytical software that are key in her field. This year she is following dedicated courses that focus on advanced programming through the high-performance computer cluster that her university makes available.

7.3 Persona #3: Ben

- **Age & Profession**: a 35-year old journalist working for the newspaper De Standaard.

- **Quote**: “Objective journalism and an opinion column are about as similar as the Bible and Playboy magazine.”

- **Short description**: Ben lives with his girlfriend in the countryside. He is passionate about everything related to cars and spends his free time fixing up old-timers with his friends.

- **Goal for using social media archive**: In collaboration with other news media outlets, De Standaard is launching a big public project on 'fake news'. One of the first tasks for Ben is to collect examples of fake news (posts) on social media that were posted in the past.

- **Professional situation**: Ben has been a journalist at De Standaard for about 4 years. Prior to that, he was head copywriter at a marketing and advertising agency. In the offices of De Standaard, Ben is known for his resourcefulness and for the fact that he likes to tinker with technology and software. However, Ben has never had a formal education in programming languages or techniques, nor does he have expertise in working with big data or analytical software.

7.4 Persona #4: Manou

- Age & Profession: 28-year old scientific researcher @ KBR

- Quote: "Those without heritage, history, and place are subject to exploitation, manipulation, and deception."

- Short description: Manou lives in an apartment in the capital together with her husband and newborn son. As her son is only a couple of months old, she spends quite some time caring for her newborn. Manou likes hiking in nature, visiting nature parks and loves to spend time on social media.

- Goal of using social media archive: Manou coordinates a four-year project dedicated to the improvement of the quality control process concerning the digitised heritage collections. One aspect of the project includes exploratory research by means of close reading of pertinent social media channels and accounts to see whether social media comments exist on these digitised heritage collections. In order to kick-off the collection she aims to collect as much as social media content as possible on already digitised heritage collections.

- Professional situation: Manou studied information and communication sciences (during which she picked up some basic digital literacy skills) and started working as an intern at the national library. After her internship, she was offered a research position at the

national library where she helps to develop the policy for long-term preservation of archived web materials and where she works on a plan for promoting the collections in the library.

7.5 Persona #5: Jan

- **Age & Profession**: 68-year old former team coordinator at a car manufacturer, now retired.

- **Quote**: "Aging is an extraordinary process where you become the person you always should have been."

- **Short description**: Jan lives in a small provincial town where he moved to after he retired. During his professional career, he worked a lot in China and he got really fascinated with this huge country. This resulted in him becoming – on a voluntary basis - the chairman of the Belgian China-Belgium club, an organization for which he does a lot of voluntary work.

- **Goal of using social media archive**: Jan regularly writes blog posts for the website of the Belgian China-Belgium club. For one of these blog posts, he would like to elaborate on how China is described and framed in Belgium, in particular on how China is portrayed on Belgian social media and how this discourse evolved over time. In order to do this, he needs all Belgian social media content about China he can get his hands on.

- **Professional situation**: Jan used to be a team leader at a car manufacturer. For this job, he had to regularly travel to China. Jan has no programming expertise, nor is he acquainted with big datasets or analytical software.