Towards a sustainable social media archiving strategy for Belgium

WP3 Report

Quality control of harvested content (Task 3.2)

Editors	Peter Mechant & Eveline Vlassenroot Fien Messens & Friedel Geeraert Sally Chambers
Responsible partners	MICT (UGent) KBR Ghent-CDH (Ugent)
Version	1.0
How to cite this?	*P. Mechant et al., BESOCIAL: Quality control of harvested content (Task 3.2), March 2022.*

Table of Contents

1. Introduction

T3.2 wants to assess the quality of the harvested content in BESOCIAL. Insights from this assessment can feed the technological development and optimisation of the harvesting process. Quality assessment or control refers to the evaluation of harvested web resources while determining whether certain quality standards are attained. The British Library uses four aspects to define quality: (i) completeness of capture, (ii) intellectual content (whether the intellectual content can be 'replayed'), behavior (whether the harvested copy can be replayed including the behavior present on the live site), and (iv) appearance or the look and feel of a website.

(Information) quality assessment or quality control (QC) is a broad field with various methodologies being used. For example, some research looks at the notion of Website Archivability (WA) and tries to capture or diagnose whether a website has the potentiality to be archived with completeness and accuracy *a priori* (e.g. see the 'Credible Live Evaluation of Archive Readiness' method[1]) as this appreciation of the archivability provides archivists with a valuable tool when assessing the possibilities of archiving material.

As the quality of information online is highly variable, other researchers to this end study users carrying out specific tasks on the Web. This has demonstrated that information quality assessment is composed of four components: credibility of content, credibility of site, predictive relevance and veracity assessment (Fink-Shamit et al., 2008)[2]. Librarians and other information professionals also use other methods or procedures to guarantee quality, substituting QC by using guidelines or fact checking[3], or by using a contextual or common sense approach. Still other approaches involve computational and automatic QC using for example neural networks that are trained on visual differences between the web page during archiving and reproduction[4].

In order to evaluate the quality of the harvested content - in specific the COVID19 related content that was harvested during BESOCIALs' 'mini-pilot' - we will adopt a pragmatic and qualitative approach that starts from the different persona that were developed in WP 2 (Task 2.2 - Analysis of user requirements). 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

[1] Banos, V., Kim, Y., Ross, S., & Manolopoulos, Y. (2013). CLEAR: a credible method to evaluate website archivability.

[2] Fink-Shamit, N., & Bar-Ilan, J. (2008). Information quality assessment on the web–an expression of behaviour. *Information Research*, *13*(4), 13-4.

[3] See e.g. https://support.archive-it.org/hc/en-us/articles/208333833-Quality-Assurance-Overview or https://www.data-archive.ac.uk/managing-data/digital-curation-and-data-publishing/quality-control/

[4] See e.g. Kiesel, J., Kneist, F., Alshomary, M., Stein, B., Hagen, M., & Potthast, M. (2018). Reproducible web corpora: interactive archiving with automatic quality assessment. *Journal of Data and Information Quality (JDIQ)*, *10*(4), 1-25.

not exactly match a specific person or comprehensively describe the full diversity of a group. Our QC using persona's can thus be considered to be a 'common sense approach' to QC as the knowledge accumulated during development of the personas will allow us to assess the quality of information.

Based on previously developed persona in the field of (web) archiving (see T2.2), and taking group discussions and the results of the PROMISE project into account, five persona were created (a full description of these 5 personas is provided in the Annex of T2.2):

- Bart - postdoc 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

In order to reduce the complexity of evaluating the harvested content from 5 different perspectives, it was decided early on in the task, to reduce these to 3 personas for the purpose of the quality assessment, namely: Febe, Manou and Jan. For the same reason, the persona description, and in specific the research question for each persona, was revised to make them applicable for the content of the BESOCIALs' harvesting 'mini-pilot'.

Next to this persona-driven research approach, this document also takes a more computational approach to assess the usability and quality of the content of the BESOCIALs' harvesting 'mini-pilot', see section 2.2.

Finally, we used the benchmark of datasets for computationally-driven research developed by Candela et al. (2021)[5], see section 2.3.

[5] Candela, G., Sáez, M. D., Escobar, P., & Marco-Such, M. (2021). A benchmark of Spanish language datasets for computationally driven research. Journal of Information Science.

2. Quality assessment methodology

This section first describes the pragmatic and qualitative approach we took to evaluate the harvested content, see section 2.1. This qualitative research approach has the advantage that it offers plenty of opportunities for exploration and discovery, for understanding the context and depth of the subject and for creating a framework of interpretation. In specific we will use the perspective of 3 personas as different frameworks of interpretation that guide the quality assessment.

In section 2.1 these 3 (revised) personas are outlined and their research questions, which will be the first and foremost interpretative perspective for the quality assessment, are highlighted. Next, a more computational approach, using Tableau and Jupyter Notebooks, is described, see section 2.2. This chapter ends with a description of the benchmark of datasets for computationally-driven research developed by Candela et al. (2021) that will be used to assess the usability and quality of the content of the BESOCIALs' harvesting 'mini pilot, see section 2.3.

2.1 Quality assessment from the perspective of 3 personas

2.1.1 Febe

Febe is a 24-year old PhD-student in computational linguistics. She lives with her girlfriend in a small studio apartment not far from university. 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. 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.

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. Given the European context of the project, Febe needs to analyse and develop the methodology for several languages. As a first test-case she has picked the COVID19-collection harvested during BESOCIAL.

2.1.2 Manou

Manou is a 28-year old cultural heritage researcher @ KBR. She 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. She 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.

Manou recently attended a WARCnet meeting of web archiving curators and researchers who work with archived web and social media data. Niels Brügger, a prominent researcher and Professor in Media Studies at Aarhus University mentioned that he is working on analysing how (the reaction to) the COVID-19 crisis differed from other pandemics such as the Spanish flu and ebola outbreaks. Manou wants to check if replicating this kind of research would be possible using the collections of KBR.

2.1.3 Jan

Jan is a 68-year old former team coordinator at a car manufacturer, now retired. He 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. Jan has no programming expertise, nor is he acquainted with big datasets or analytical software.

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 depicted on Belgian social media in relation to COVID19 and how this discourse evolved over time. In order to do this, he starts off with the COVID19-collection harvested during BESOCIAL.

As mentioned before, we will use a ‘common sense approach’ to QC as the knowledge accumulated during development of the personas will allow us to assess the quality of information. The quality assessment procedure will also involve a SWOT-analysis during which several researchers reflect on the strengths, weaknesses, opportunities and threats that the COVID19-collection harvested during BESOCIAL offers for tackling the following research questions:

- What is a suitable methodology for sentiment analysis and subjectivity detection through the deep semantic analysis of text published via social media channels? (Febe)

- How does (the reaction to) the COVID-19 crisis differ from other pandemics such as the Spanish flu and ebola outbreaks? (Manou)
- How is China described and framed in Belgium, in particular; how is China depicted on Belgian social media in relation to COVID19 and how did this discourse evolve over time? (Jan)

The evaluation of the COVID19-collection based on these persona-driven research questions will be further supported by in-depth interviews with respondents who's features align with the previously defined persona. The following interviews were conducted to support the narratives of Febe, Manou and Jan:

- Febe: two interviews with researchers at Language and Translation Technology Team, a UGent research group which conducts fundamental and applied research on different aspects of Natural Language Processing (NLP) from a corpus-based inductive perspective (21/01/2022 & 24/01/2022)
- Manou: two informal interviews with researchers at KBR active in the field of web and social media archiving (16/12/2021 & 09/02/2022)
- Jan: two interviews with a retired project manager (11/12/2021 & 09/01/2022)

2.1.4 SWOT analysis technique

The SWOT analysis technique (see Figure 1) is often employed in business analysis for identifying factors influencing a company's position in the market. However, the SWOT framework can also provide significant value outside of the business domain as it's essential aim is to assess internal (Strengths and Weaknesses) and external (Opportunities and Threats) elements of the studied subject c.q. the usefulness and quality of the COVID19-collection harvested during BESOCIAL for a diversity of people (in specific the personas Febe, Manou and Jan).

It should be noted that a SWOT analysis is not as straightforward as it seems. Categorization is a subjective process and depending on how they are framed, strengths can also be weaknesses (and vice versa) while opportunities can also be threats. Furthermore, the boundary between internal and external can be vague and subject to interpretation and discussion. Despite these shortcomings, we think a SWOT is a viable framework to structure this quality assessment and that the results will provide a starting point for further discussion (and optimisation of the harvesting process).

'Strengths' refer to attributes of the COVID19-collection that are helpful for tackling the research questions, 'weaknesses' refer to attributes of the COVID19-collection that are not present or detrimental for solving or tackling the research questions, 'opportunities' summarize the external conditions helpful for working on the research questions and 'threats' point to the external conditions that could threaten or hinder working on the research questions and using the BESOCIALs' COVID19-collection.

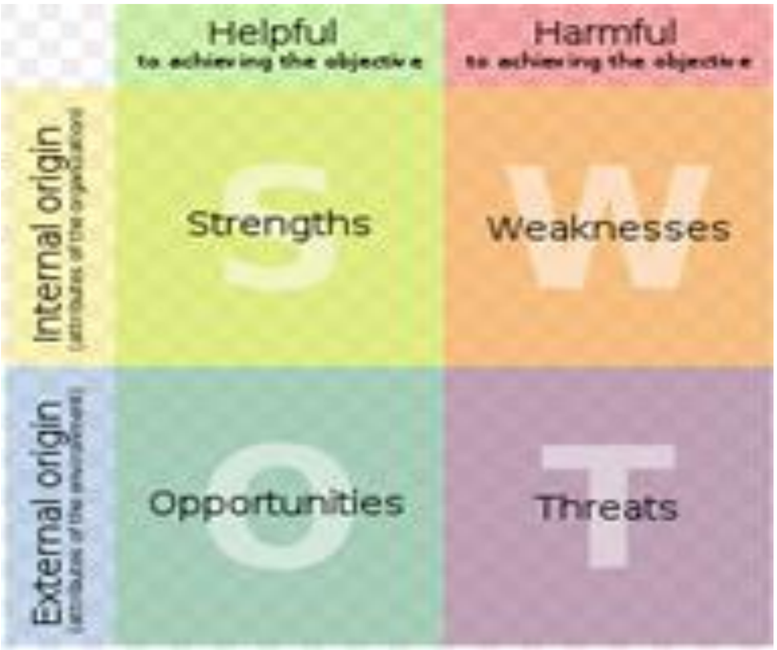

Helpful
to achieving the objective
Harmful
to achieving the objective
Internal origin
(attributes of the organization)
External origin
(attributes of the environment)
S
Strengths
W
Weaknesses
O
Opportunities
T
Threats

Figure 1: Abstract representation of SWOT components

2.2 Quality assessment using Computational Methods

Besides these persona-driven research questions that will be explored in a qualitative manner, we decided to add these more generic research questions - to be investigated using a computational approach - to the assessment procedure:

- Percentage of unique twitter users that disclose location (in tweet)
- Percentage of unique tweets that contain named entities that appear in dbpedia (relates to the percentage of tweets about or by a ‘private public person’)
- Discover tweets that have been deleted and investigate the reason behind this deletion (ref. Covid19 vaccines prices tweet)
- Determine if any given tweet or account is still live? (rehydration)
- Determine if all (or 3200 max) tweets of a given account have been archived
- Determine if all short-urls get resolved to the (long) original URL

2.3 Quality assessment using Candela et al.'s (2021) benchmark

Recent work by Candela et al. (2021) proposes a methodology to select datasets for computationally-driven research. Although this work is applied to Spanish text corpora, we have tried to extend the work to a broader context in order to use the methodology as a benchmark for the COVID19 related content that was harvested during BESOCIALs' 'mini-pilot'.

Given that benchmarks provide an opportunity for comparing and assessing the 'performance' of databases with those regarded as the best and given that it allows for the identification of opportunities for improvement, the benchmark by Candela et al. (2021) is used in this context to assess to what extent the BESOCIALs' COVID19 related content is amenable for computationally driven research. Each benchmark's criteria scores according to a criterion that consists of a function, with values ranging from 1-0. Candela et al. (2021) propose the criteria:

- Licensing (from very permissive with none or few obligations and known as open, to very restrictive or closed)
- Accuracy (extent to which data are correct, reliable, and certified free of error)
- Provenance (description of the creation process and the derived data)
- Language (availability of languages in which we are interested)
- Permanent identifier (e.g. by the assignment of a DOI)
- Prototypes and documentation (providing prototypes & examples of use in addition to documentation)
- Formats (variety of formats)
- Terms of use and code of conduct (presence of terms of use to the datasets)
- Technical aspects (in terms of access possibilities)

For a full description about how the criteria above are defined, see Candela et al. (2021).

3. Results

In this section we describe the results of our quality assessment procedure on the COVID19 related content that was harvested during BESOCIALs' 'mini-pilot'. First the results of the SWOT analysis technique using the perspectives of the persona Febe, Manou and Jan are discussed. Next, the results of the exploratory computational quality control are briefly presented.

3.1 SWOT analysis technique using the perspectives of the persona's

3.1.1 Febe

3.1.1.1 Febe's narrative

Febe's goal is to find a suitable methodology for sentiment analysis and subjectivity detection through the deep semantic analysis of text published via social media channels. In contrast to the other to persona (Manou and Jan), which will be discussed below, Febe has strong computational and coding skills which she acquired in her Master of Arts in 'Technology for Translation and Interpreting' during which she received courses on Python programming for example.

Her strong computational and coding skills immediately become clear; after unzipping the .json-file it only takes her a couple of seconds to **grasp the content and structure of the file by using 'less'**; a terminal pager program on Unix, Windows, and Unix-like systems used to view the contents of a text file one screen at a time. It is similar to 'more' , but has the extended capability of allowing both forward and backward navigation through the file.

Once it is apparent to Febe that the file contains a standard-dump of Twitter tweets in .json-format, she quickly modifies a python-script she has used earlier, **to import the data into the python programming environment**. Python is an interpreted high-level general-purpose programming language. Its design philosophy emphasizes code readability with its use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. In this case using functions such as **'json load' or 'json dump' in combination with regular expressions enable Febe very fine grained control about which content she wants filtered or not.**

After ingesting this content in the Python environment Febe has all her standard scripts and tools at her disposal to further drill-down into the content. She can create sentiment classifiers, and machine and deep learning routines using PyTorch for example, which she can 'train' using a part of the dataset (training data). Next, she can test the code she has generated on another, independent part of the data set (test data) in order to evaluate the efficacy of her scripts. Python also has various libraries (e.g. MatPlotLib) that enable Febe to easily create visualizations of (parts of) the dataset.

3.1.1.2 'Strengths'

Febe mentions that she, and some of her colleagues, regularly use the Twitter API to collect content to use in their research or that she reuses content that was collected by others. In that context, the json-dump of the COVID19-collection harvested during BESOCIAL feels very 'natural' for her as she has the tools and expertise to process, clean and analyse such data. Febe also mentions formats such as .txt or .csv to be amongst the file formats most often used.

3.1.1.3 'Weaknesses'

Febe points out that not much information is available about the harvesting process nor about how the seed lists were constructed. She remarks that the usefulness of the COVID19-collection for her personally is strongly weakened by the absence of such information as she needs a maximum degree of transparency in order to assess e.g. the potential biases present in the collection. She also regrets the fact that no basic collection descriptive (e.g. amount of tweets, retweets, … but also top hashtags, embedded media content etc.) is available.

3.1.1.4 'Opportunities'

Febe distinguishes two main opportunities in the COVID19-collection harvested during BESOCIAL namely the 'domain' specific and the multilingual character of the dataset. The dataset is 'domain' specific in that it focuses on communication about a specific topic (covid19) on a specific communication channel/platform (Twitter). This specificity proves to be beneficial for (computational) linguistics and NLP researchers as it offers a well-defined scope and focus (in terms of lexicon, vocabulary, formality of speech used in the data collection) to delineate research activities. Related to this is the fact that the multilingual character of the COVID19-collection enables (computational) linguistics and NLP researchers to consider different languages within this well-defined domain specific language use.

3.1.1.5 'Threats'

Polled for potential threats, i.e. external conditions that could threaten or hinder Febe working on her research questions and using the BESOCIALs' COVID19-collection, Febe again points to the lack of descriptive metadata for the COVID19-collection harvested during BESOCIAL. She experiences this as a 'threat' because she finds it difficult to assess potential biases in the collection which might potentially diminish the validity of her research.

3.1.2 Manou

3.1.2.1 Manou's narrative

Manou recently attended a WARCnet meeting of web archiving curators and researchers who work with archived web and social media data. Niels Brügger, a prominent researcher and Professor in Media Studies at Aarhus University mentioned that he is working on analysing how (the reaction to) the COVID-19 crisis differed from other pandemics such as the Spanish flu and ebola outbreaks. Manou wants to check if replicating this kind of research would be possible using the collections of KBR. This is a test case for Manou and she wants to do a quick **exploratory review of the existing data in KBR's collection** to see whether she can create a corpus that is substantial enough to answer her research question.

She starts by focusing on the **social media corpus on COVID-19**. The files were made available for download and as the files are large, it takes over an hour to download and unzip them. She has to use the software editor Atom to open the .md file on her personal computer as the laptop from her work does not have the necessary software. She has already created a ticket to request Notepad++ to be installed on her laptop, but the request has been pending for over two weeks. She opens the read.me file accompanying the folders 'accounts', 'dutch-hashtags', 'french-hashtags' and 'mix-language-hashtags' but she doesn't see immediately which accounts or hashtags have been archived. She would like to see if the social media accounts of any newspapers were archived.

She decides to take a look at the data in the folders and starts with the 'dutch-hashtags'. She wonders why there are so **many read.me files** instead of a single general one. In the file with name 3e92876a5a5e4169986185013285d98d-README, Manou does not understand what "Change to Twitter Dutch corona hashtags (collection) on Jan. 27, 2021, 2:08:53 p.m. CET by sven: * is_on: "False" changed to "True"" means and regrets that no link to a file with additional information about how to interpret these READ ME files is provided as she is not familiar with SocialFeedManager as a tool.

Manou opens the **two Excel files containing the Dutch hashtags**. One of them (c3ee59b14a5446c0abc28cc3328ef080_001) is structured in the sense that the data is subdivided into different columns, while the other file (4feabc2be79849699d455543e7ab1006_001) needs to be divided into columns. Manou uses the 'Text to Column' option to **organise the data** in the second file into columns. Manou feels overwhelmed by the sheer amount of columns and data, but after scrolling through the column titles, decides that the column 'text' would be the most useful to start with given that it contains the text of the tweets/retweets. A file detailing the content of each column would have been useful because Manou cannot correctly interpret the meaning of columns such as 'possibly_sensitive'. A quick Google search leads her to the Twitter Developer Platform where she discovers that the field is a Boolean and indicates that the URL contained in the tweet may contain content or media identified as sensitive content. She scrolls through the first lines of the Excel files and notices that there are a lot of RT or retweets that repeat the same statement. She notes that the **duplicates** will have to be removed in a data cleaning phase and decides that, if she will move forward with this research question, she will have to use a **data cleaning tool** like **OpenRefine** to remove these duplicates from the dataset. Manou tries to open the tool OpenRefine that already has been installed on her laptop. The tool won't respond. She tries to download a more recent version, but has to ask the ICT helpdesk for the admin password. Because of this, Manou decides to leave the idea of using OpenRefine and focuses on using Excel as a cleaning tool.

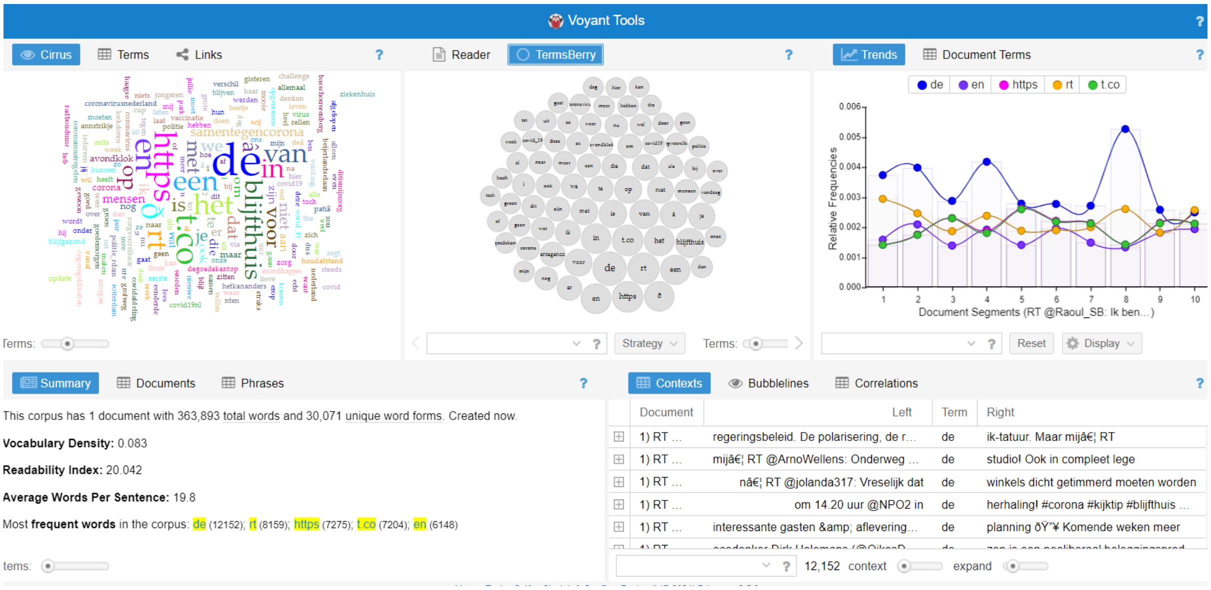
Upon further reflection, she sees that there is a column named tweet_type with values 'original', 'retweet'', quote', 'reply', 'retweet' and 'blank'. She decides against using OpenRefine and using a filter on this column instead to filter out the retweets. She notices that there is a **lot of content pertaining to emotional reactions to the COVID-19 pandemic** with tweets stating: 'Ik vind er niks aan zo, met die helicopters boven ons huis. #avondklok #avondklokrellen #osdorp #blijfthuis #alsjevanjestadhoudthoudjejestadheel'. Manou also notices that a lot of tweets are posted by residents in The Netherlands such as 'Lieve Zeeuwen, Middelburgers,Vlissingers,Goesenaren..#bluufthuuuuss Lieve ouders van alle kinderen die naar buiten willen..laat ze niet gaan...#blijfthuis #geenrellenaub #doeenslief #denkaaneenander'. She realises that if she wants to use this corpus for her research it will require thorough **data cleaning** to only obtain Belgian data. She wonders if she has enough time to do so and whether in that respect, the collection of the accounts would be a better choice as the curators would in all likelihood have been better able to determine which accounts were linked to Belgian organisations etc. than which Dutch hashtags are used in Belgium alone.

Manou opens the **Excel file of the accounts** (059645541e2e4e1cbd719f2eacd7c949_001) and wonders first why there is only one excel file as for the Dutch hashtag collection there were two. She tries to open the document, but the file is very large and her laptop takes time to open the file which leaves Manou frustrated. Once the file is opened, she tries to use the 'Text to Column' functionality in Excel to structure the data, but her computer becomes unresponsive due to the size of the file. After a while her computer pulls through and allows the data to be split into columns. She is asked by Excel if she wants to continue to save the data as CSV and she clicks yes. After scrolling through the data she notices that something is wrong with the text encoding as the accented letters in French for example are displayed as symbols.

Manou then moves on to the BelgicaPress **and** BelgicaPeriodicals platforms. As the former contains 114 Belgian newspapers (1814-1970), she hopes that she will be able to find information about the Spanish flu in 1918. She searches for 'Spaanse griep' and 'grippe espagnole' and finds relevant articles in the Newspapers 'De Schelde', 'Ons Vaderland', 'Nieuwe Gazet' and 'Le Soir'. In the periodicals collection, she identifies content in the ‘Hebdo Pourquoi Pas’ and ‘Libra Illustré’. No Dutch-language periodicals were found. Manou realises that, in order to have a newspaper counterpart that is significant enough to balance the social media corpus, she will have to dig deeper and consult the paper newspaper and periodical collections at KBR to create a bigger corpus. This, however, would take her a lot of time and she isn’t sure she will have enough time to do so.

Phase 3 quick overview

Manou tries to get a quick overview of what the text within the Dutch hashtags collection is about. She uses the Voyant Tools **word cloud** tool that she learned during her Bachelor years at the university (see Figure 2). It isn’t the most representative output, but it can sure give her an idea of how the text looks like. The tool runs very slow and isn’t the most efficient way to upload a corpus.



This corpus has 1 document with 363,893 total words and 30,071 unique word forms. Created now.

Vocabulary Density: 0.083

Readability Index: 20.042

Average Words Per Sentence: 19.8

Most **frequent words** in the corpus: de (12152); rt (8159); https (7275); t.co (7204); en (6148)

Document	Left	Term	Right
1) RT ...	regeringsbeleid. De polarisering, de r...	de	ik-tatuur. Maar mijâ€¦ RT
1) RT ...	mijâ€¦ RT @ArnoWellens: Onderweg ...	de	studio! Ook in compleet lege
1) RT ...	nâ€¦ RT @jolanda317: Vreselijk dat	de	winkels dicht getimmerd moeten worden
1) RT ...	om 14.20 uur @NPO2 in	de	herhaling! #corona #kijktip #blijfthuis ...
1) RT ...	interessante gasten &amp; aflevering...	de	planning ðŸ”¥ Komende weken meer

12,152

Figure 2: screenshot Voyant tool

Manou decides to delete certain stop words, but the program isn’t responsive. She comes to the conclusion that Voyant tools and the creation of a word cloud wasn’t the best choice to get a better understanding of the data.

Phase 4 research

She decides to call a friend who recently finished a master in Digital Humanities at the KULeuven to discuss which tools would be best suited to the data and her research questions. They also discuss possibilities to do **sentiment analysis**. Her DH friend advised her to do a first step of cleaning the data.

She installs **OpenRefine** on her personal computer and cleans the data by removing duplicates from the **archived social media posts** and posts that are not related to Belgium. She uses the Dutch guideline on OpenRefine available on the CEST website. A year ago, she also took a workshop at meemoo, which taught her the basics of working with data cleaning tools, such as Open Refine. After the cleaning procedure, she then scores the comments in the **cleaned dataset** in Excel on a **5-point scale** ranging from very negative to very positive in order to gain more insight into the predominant sentiment.

She then compiles the reactions to and mentions of the Spanish flu she found in **BelgicaPress** and **BelgicaPeriodicals** in Excel and repeats the same exercise using the 5-point scale.

She complements the mentions of the Spanish flu in digitised newspapers by compiling a corpus of mentions of the Spanish flu in the **paper newspaper collections** of KBR. KBR's catalog contains 150 records for newspapers published in 1918. She therefore decides to select 4 newspapers, two published in Dutch and two published in French that were national in scope (Belgisch Dagblad, Vrij België, L'Indépendance belge and Notre Belgique quotidien). She copies relevant pieces of text into an Excel document and scores the corpus using the 5-point scale. Compiling the corpus based on the paper collections is labour-intense and she regrets that not all newspapers in KBR's collections have been digitised and made full-text searchable.

Having compiled and scored her corpus, Manou draws **conclusions** about the similarities and differences between the reactions about COVID-19 and the Spanish flu in 1918. She also lists the **limitations** of the study and describes the **methodology** and reasoning behind the study.

Phase 5 visualizing

Her DH friend that she consulted a few days ago also suggested visualizing the data. Manou decides to write to her mentor from her internship at Kantar that she did during her masters. Manou remembers she has a Phd in computer sciences focused on visualizing big data. **Tableau** could be a feasible tool, recommends dr. Jansens. Manou is in doubt whether she should be using the Desktop version or the online version. After all, she chooses the free trial Online option.

After being a bit overwhelmed by all the functionalities of this visualization tool she decides to upload the 5 point scale excel file she created manually. Manou creates simple bar charts of the outcome of the sentiment analysis by creating simple graphs based on the number of posts attributed to each category of the 5-point scale.

Phase 6 dissemination

Manou decides to share her findings with her colleagues at KBR during one of the '**Research lunches**' that are organised once a month. Furthermore, she submits an abstract for a short paper for the **DH Benelux conference** as it is a good illustration of how KBR's diverse collections can be used.

Manou hopes to present her findings of this small research project in person in Luxembourg. There she plans to network and receive constructive and interesting feedback. The ultimate goal of submitting an abstract to this international conference will be a publication in the **DH Benelux Journal**. The publication will not be ranked (not A1 or A2), but will still be a good approach to promote her research at KBR.

3.1.2.2 'Strengths'

- Born-digital data made available as downloadable data is a big plus
- Manou's position at KBR also allows her to speak to the colleagues who have worked on the collections (BESOCIAL, BelgicaPress and BelgicaPeriodicals) and others who have ample expertise in working with born-digital data which allows her to acquire additional contextual information.
- Online access to BelgicaPress and BelgicaPeriodicals is a very powerful tool to quickly get an overview of what kind of data is available in KBR's digitised newspaper and periodicals collection. The full-text search is very helpful.

3.1.2.3 'Weaknesses'

- The data needs cleaning (duplicates, non-relevant content such as tweets from Dutch citizens in the Dutch hashtag collections)
- No 'customisation' of the social media dataset can be done. Researchers cannot for example specify the period that interests them, for example 'all tweets in the dataset published in November 2021'.
- The documentation needs to be more clear (subdivisions hashtags and accounts)
- Corpus is too big for close reading only
- If Manou wants to install certain software on her laptop to clean and/or visualize the data she is dependent on the helpdesk at KBR. The request to install tools takes some time.
- BelgicaPress and BelgicaPeriodicals do not allow to download the corpus as data. In Manou's case the corpus is rather small, so she can compile it manually, but for larger datasets, export options should be provided to researchers.

3.1.2.4 ‘Opportunities’

- Dissemination opportunities:
    - The corpus can be promoted at universities so that interested students can use it for their bachelor paper / master thesis
    - The social media corpus can also be used as a subject for internships at KBR
    - Conferences such as the IIPC WAC and RESAW conference can provide interesting platforms to promote the collections on an international level.

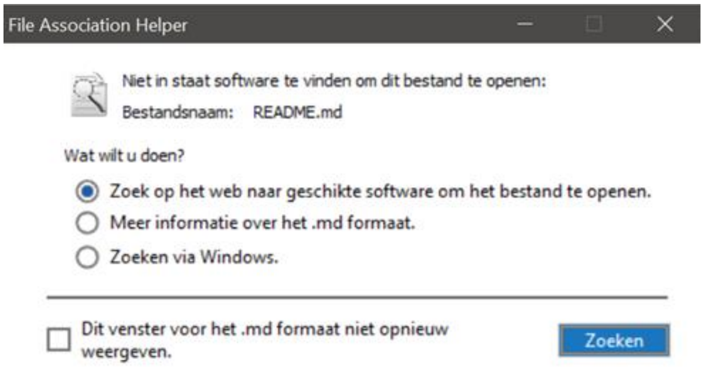
3.1.2.5 ‘Threats’

- The public who would use this data, will not have access to Manou’s colleagues. The development of additional information about the data and concrete examples of how the data can be used to do research is needed.
- Storage capacity

3.1.3 Jan

3.1.3.1 Jan’s narrative

In order to answer Jan’s curiosity on how China is depicted on Belgian social media in relation to COVID19 and how this discourse evolved over time, he starts by **unzipping** the compressed folder he **downloaded** from the data source provider. Jan notices the README-file in the uncompressed folder and decides to use that file to start his exploration of the archive as he believes it will provide him with useful information. However, when he tries to **open the README-file** he notices the file-extension ‘.md’ which he is not familiar with. More importantly he notices – by means of a pop up – that he has no software installed to open this file (see Figure 3).



Niet in staat software te vinden om dit bestand te openen:
Bestandsnaam: README.md

Wat wilt u doen?

● Zoek op het web naar geschikte software om het bestand te openen.
○ Meer informatie over het .md formaat.
○ Zoeken via Windows.

□ Dit venster voor het .md formaat niet opnieuw weergeven.

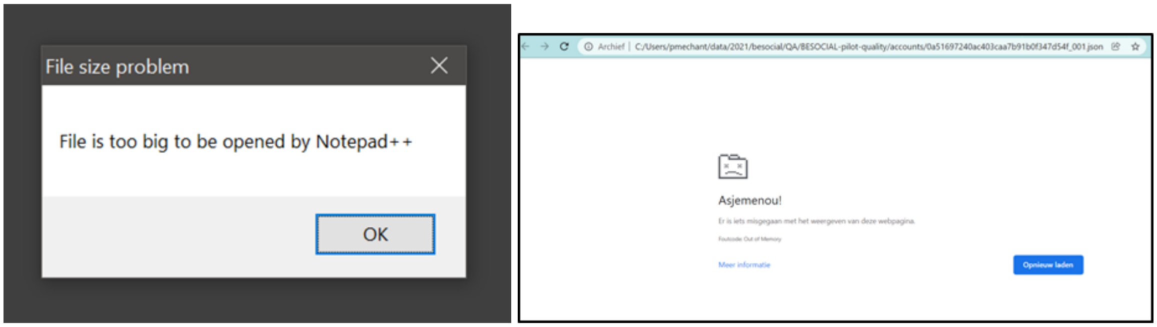
Figure 3: Jan is unable to open the README.md file

Using Google, Jan learns that this type of file can be opened with any text editor, including: Microsoft Notepad (Windows), Apple TextEdit (Mac) or Vim (Linux, Mac). Jan works with a Windows OS, so after selecting 'Search using Windows' in the pop-up he succeeds **in linking Notepad to the .md-extension**. Though **Notepad seems to crash on the file**, he succeeds in opening the file using **Notepad++**. **Reading the file** Jan **learns some basic information**, namely that the folder he downloaded contains Social Feed Manager exports from the BESOCIAL mini pilot and that the data from the three hashtag collections was exported on November 16 and the data from the account collection on November 26. Jan wonders if this refers to November of this year or earlier years and regrets not finding a hyperlink to a webpage providing more information.

In the uncompressed folder Jan notices different subfolders. Jan decides to focus on **'accounts' and opens this folder**. At first, Jan is a bit confused; the folder contains 3 different README-files with very **cryptic and complicated filenames**. He decides to have a look at the 3 files and decides that these describe the same collection as they all mention the **same collection id** (1263b5e366e74bfb9a643157bb457af6). Jan thus decides to only read the file from the most recent export, namely the README-file created on Nov. 26, 2021, 3:50:52 p.m. CET. He then notices that the filename '059645541e2e4e1cbd719f2eacd7c949-README.txt' corresponds with another file in the folder. Jan **recognizes the Excel logo** and is assured that this file will contain the data described in the README-file.

Still, he wonders why the file with the extension .json is even larger in file size. Jan tries to open the file '0a51697240ac403caa7b91b0f347d54f_001.json' but encounters the same difficulties as when trying to open the .md-file. Thus Jan tries to open the .json-file using 'Notepad++'. However, without success; the application returns an error pop-up stating 'File is **too big to be opened** by Notepad++' (see Figure 4). After some Googling and using the Windows-wizard to search for a suitable software application for opening files with the json-extension, Jan succeeds in viewing the content of the json-file through his internet browser. The 'webpage' with the content however seems to be **loading indeterminately** and does not succeed in showing all the content or simply crashes (see Figure 4).

Giving up on this file, Jan tries to open and explore the text file in the folder named 'dae3fce02fc340ffac7e9ec37934d377_001.txt'. In his text editor he notices that this file contains a long list of identifiers that are made up out of 19 numbers. As he is familiar with Twitter, Jan recognizes this format as a tweet-id (https://developer.twitter.com/en/docs/twitter-ids). He puts his hunch to the test using a simple URL-hack; first he navigates to Twitter and selects the URL of a random tweet from a Twitter account Jan is following on his timeline (e.g. https://twitter.com/kbrbe/status/1468947413410983938). He then replaces the identifier in this URL with an id found in the text file with identifiers. This is successful as the new link brings Jan to a tweet from January about Covid19 which he supposes was archived.



File size problem

File is too big to be opened by Notepad++

Archief | C:/Users/pmechant/data/2021/besocial/QA/BESOCIAL-pilot-quality/accounts/0a51697240ac403caa7b91b0f347d54f_001.json

Asjemenou!

Er is iets misgegaan met het weergeven van deze webpagina.

Foutcode: Out of Memory

Figure 4: Jan encounters the size limits of Notepad++ and his internet browser

From the README-file in the 'accounts' folder Jan learns that **the export** was created with Social Feed Manager and that it contains all seeds. His basic knowledge of Twitter also helps him in deducing the metadata field "Export type: twitter_user_timeline" and the field "Deduplicate: Yes". Under the heading '**Collection information**' Jan finds some useful information, namely that the collection contains all Twitter accounts from the project's selection which were harvested once per day. Jan also notices the 'seeds' or accounts that were archived. He regrets that this information is not available in a more structured format and copies the text to a word-file to format the text in order to get a clearer overview of the seeds that were used to create this collection. Scrolling down through the **change-log** Jan notices that an account name was renamed in March and he deduces that the archival **collection was started on January 25th 2021**. Reading the description 'schedule_minutes: "blank" changed to "1440"' he confirms that 1440 refers to minutes as the harvesting was done on a daily basis.

When opening the **.csv-file** Jan notices that the data is **not structured in columns** but only in rows, with commas separating different data fields. This makes it very difficult to have a clear overview of what the file actually contains. In order to ascertain the presence of tweets containing information about China, Jan does a **search query (using Ctrl+F)** on 'China', causing his laptop to almost stall. He notices that 874 cells are found containing 'China'. Happy with this result he decides to contact his son and **ask him for help** in processing a .csv-file.

Jan's son demonstrates to his father how he needs to start from a blank Excel-sheet and use the function 'Data > **Import from text or csv**' to read the .csv-file correctly in Excel. This proves rather easy but Jan almost immediately notices that **some text is not formatted correctly**, e.g. ""Nous devons Ãªtre trÃ̈s inquiets". Ignoring this Jan decides to first **reduce the data set** to the information he is interested in.

For this he applies a **text-filter** on the 'text-column (using contains 'chin*'). This results in 2643 records found. Jan decides to copy these records to another tab. Next, he deactivates the current filter and executes the same text-filter on the column 'hashtags' resulting in 101 records that are copied and then pasted in the other tab underneath the previous results. Jan realizes that the new tab he copied data to might have some duplicates (i.e. tweets containing 'chin*' in both text and hashtag-field). Thus he **deduplicates** data in this tab using the built-in Excel-functionality.

In order to have a better overview of what this 'China'-sub collection that he created actually contains, Jan decides to calculate some **descriptive metrics**. In order to determine the number of unique accounts present in the selected dataset, Jan again used deduplication; this shows that 65 unique accounts have sent the tweets Jan selected. Next, Jan uses this information to execute a 'count.if' instruction for each of these 65 accounts, resulting in a table and figure that shows the amount of tweets for each unique account. In this way Jan learns that almost 60% of all tweets originated from 6 accounts (see Figure 5).

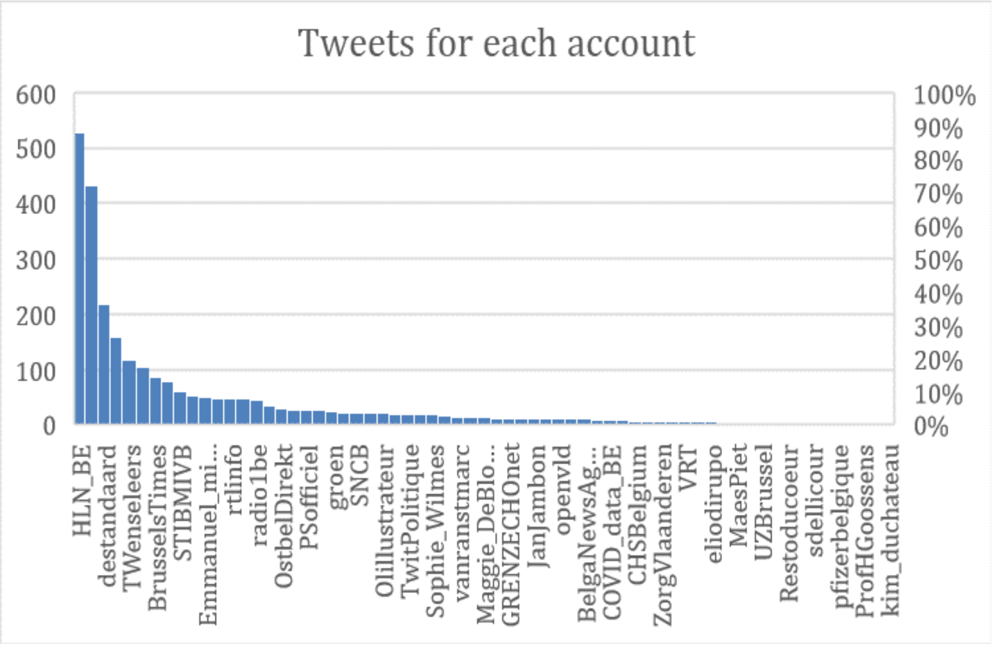


Figure 5: Jan uses Excel to determine number of tweets for each account

In a similar fashion, using basic Excel functions, Jan is able to calculate other simple metrics about the 'China'-sub collection he created, e.g., he can calculate the percentage 'original' tweets versus 'retweets', the percentage of URLs in an average tweet, the amount of Dutch of French tweets or the average amount of retweets or favorites. Also, using the data in the 'hashtags'-column Jan creates a word cloud using a freely available online word cloud generator (e.g., https://www.wordclouds.com/). However, when trying to visualize the tweets in a timeline, Jan encounters the intricacies of the date/time format in Excel and he gets stuck converting the 'created_at' column in a usable date-format.

Next, Jan decides to tackle the actual tweets themselves and to perform a **qualitative content analysis**. However, at this stage he notices that implementing his text-filter 'chin*' resulted in the inclusion of certain tweets that are not about China but that contain the expression 'chin*' (e.g., "Alec Baldwin omarmt echtgenoot Halyna Hutchins tijdens emotionele ontmoeting"); so he manually removes these cases. He then copies and pastes the remaining tweets into a new tab. In different columns Jan starts the coding process for each tweet using the grounded theory-approach. This coding approach consists of an iterative process during which first 'open codes' are assigned. These open codes simply identify categories and their features and are a first step in reducing the text towards meaningful codes. Next, Jan creates 'axial codes' that link and relate different categories with each other and reduce the number of codes. As a last step, Jan assigns 'selective' codes integrating codes and their relationships in one focused theory.

Jan also shows interest in conducting a **sentiment analysis** on the tweets he selected. However after spending some time online researching the possibilities he decides that this goal is out of his reach. Although he quickly finds some free and interesting looking software online (e.g. https://text2data.com/Excel) he also learns that these tools primarily target sentiment analysis of English content and do not provide support for e.g., Dutch. He discovers that the available software for Dutch sentiment analysis is still rather too technical and complex for his level of data or coding literacy (e.g., https://www.clips.uantwerpen.be/clips.bak/pages/pattern).

3.1.3.2 'Strengths'

One of the main 'strengths' or 'affordances' of the COVID19-collection proved to be the **.csv-format** in which the data was provided. Although Jan needed some help importing the file into Excel, from then on, he was able to use his rather limited knowledge of the Excel functionalities to create some descriptive metrics and visualizations about the 'China'-sub collection that he selected.

He was able to isolate the text-content of each tweet in order to process it using (manual) qualitative content analysis techniques. Providing **the full seed-list** of the COVID19-collection also enabled Jan to easily determine the ratio of unique Twitter accounts in the collection (that tweeted about China). In short, by providing Jan with a file-format that could be imported in Excel, a spreadsheet software package that is used by over 1 billion people worldwide (https://www.senacea.co.uk/post/excel-users-how-many), low data or coding literacy is required to conduct basic analytics on the data, while ensuring compatibility with other popular desktop software such as Word.

3.1.3.3 'Weaknesses'

Although working in Excel requires rather low data or coding literacy and skills, the story of Jan also unearths some weaknesses related to this approach which might be detrimental for Jan to solve or tackle the research question he is interested in. One of the main weaknesses in offering the COVID19-collection lies in the fact that **basic Excel knowledge is required** to import, to filter content and to calculate basic metrics about the collection as the README.txt-files provide little contextual information next to the seed-list. Another feature of offering the COVID19-collection (**three differently named README.txt-files**) caused quite some confusion with Jan and can be thus considered as a weakness. This also applies to the rather cryptic naming of the files in the unzipped COVID19-collection.

3.1.3.4 ‘Opportunities’

In terms of opportunities that summarize the external conditions that are helpful for working on Jan’s research question, we notice that **literacy** in terms of data and coding skills and to a broader extent, **social capital** in general, play an important role in determining to what extent Jan can resolve his research question using the BESOCIALs’ COVID19-collection (having a wide range of connections, e.g. his son, enables Jan as he is able to get work done more effectively and efficiently). From this perspective opportunities arise to make Jan’s data processing and analysis more fluent by implementing, what could be described as, ‘low-hanging fruit’ modifications to how the collection is offered, e.g., by:

- providing the data directly in Excel (xls or xlsx-format) thus mitigating the import issues Jan encountered
- offering filtering options on the BESOCIALs’ COVID19-collection prior to actually (having to) download the whole dataset
- creating and visualizing an ‘on the fly’ dashboard that displays and summarizes the main descriptive metrics of the collection
- providing (links to) tutorials and guidelines on how to process and analyse the collection using of the shelf desktop software
- offering examples and best-practices on how to use the collection

3.1.3.5 ‘Threats’

In line with the previously identified opportunities the main external conditions that threatened or hindered Jan’s work on his research questions were related to his data and coding skills and the potential support network he has (no) access to. Data literacy proves to be key here. It encompasses skills and competencies essential for meaningful and informed participation in today’s society and is vital for managing, accessing and critically analyzing data and the data-collection process. This threat, combined with the lack of extensive descriptions of the collection, let alone examples and best-practices on how to work hands-on with the collection, makes it rather difficult for Jan to resolve his research question in a fluent and timely manner.

3.2 Results exploratory computational quality control

Besides the persona-driven QC described above, we used a exploratory computational quality control focussing on determining amongst others:

- Percentage of unique twitter users that disclose location (in tweet)
- Percentage of unique tweets that contain named entities that appear in dbpedia (relates to the percentage of tweets about or by a 'private public person')
- Discover tweets that have been deleted and investigate the reason behind this deletion (ref. Covid19 vaccines prices tweet)
- Determine if any given tweet or account is still live? (rehydration)
- Determine if all (or 3200 max) tweets of a given account have been archived
- Determine if all short-urls get resolved to the (long) original URL

Below, we briefly present the results of this exercise. First we describe how we tried to tackle the above-mentioned research questions using Tableau, an interactive data visualization software tool. Next, we present the results of an analysis using Jupyter Notebooks, a web-based interactive development environment for data science and scientific computing.

3.2.1 Results exploratory computational analysis using Tableau

Below we describe how a researcher could explore the Twitter sample .csv file using the software tool Tableau. A few metrics were defined above that served as an inspiration to determine which metrics could be easily visualised using Tableau.

The first step was to prepare the .csv file for loading into Tableau. We did this by separating some columns in the delimiter, including the column for the hashtags. We also transformed the text from ASCII to Unicode (UTF-8). This was needed because we saw some of the Tweets were in French and were using accent marks (e.g. é; â; ç) which were not readable in ASCII. We performed both these steps in MS Excel.

The second step consisted of loading the clean .csv file into Tableau. We did this by setting the .csv file as the data source for our dashboard and adjusting some data types of the columns. After this, we made separate sheets in Tableau for each computational metric. We did this based on a quick exploration of the .csv file and after playing around with the several metrics in the file, for example count unique hashtags, twitter users etc. This exploration took some time because it was needed to understand the column names that were used.

As a result we visualised the following metrics:

- Occurrence of each hashtag in all the collected Tweets
- Unique amount of hashtags used by each Twitter
- Unique Twitter accounts using a certain hashtag
- Users disclosing their location
- Twitter accounts with an URL in their profile

These measures were visualised using only bar charts, other more advanced visualisations such as tree maps, packed bubbles could also be used but because of the size of the data we opted for the most basic representation. Also, the values are presented in the dashboard rather than the percentages. The results can be consulted via this link.

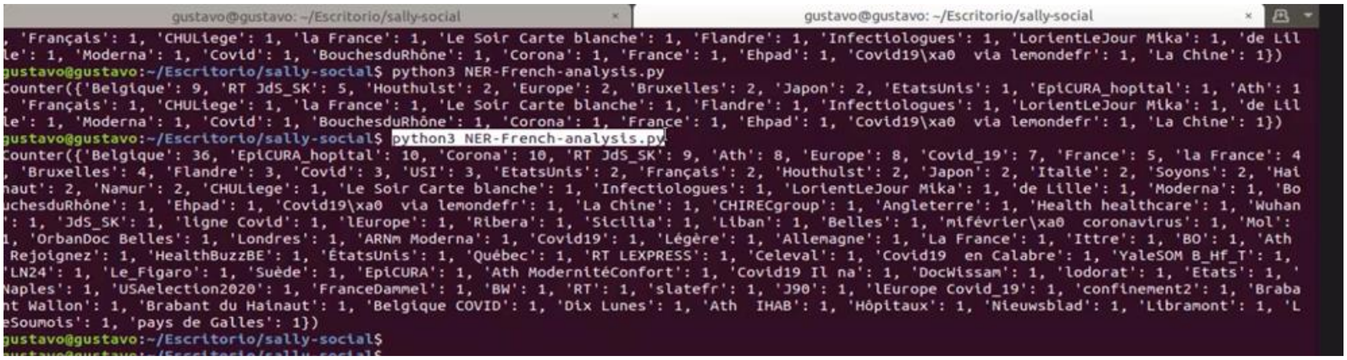
3.2.2 Results exploratory computational analysis using Juypiter Notebooks

Below, we sketch how a researcher could potentially explore the database using python and Jupyter notebooks[6]. To be specific, we demonstrate how one could analyse the bigrams in the lexicon of the BESOCIALs' mini-pilot database.

The first step involves compressing the database – in this case the .csv file with French tweets – using Pandas and reducing the file size from 250MB to 46MB. Next, the text is cleaned for punctuation, stop words etc. Using the resulting text file, a simple word count can shed some light on the actual contents of the BESOCIALs' mini-pilot database by sorting from most frequent to less frequent occurring words.

Next, a named entity recognition script (NER) can be applied to the text to extract named entities, using a French language model from Spacy NLP. Subsequently, the python script tried to assess whether the named entity referred to a specific geographical location. If so, this location was added to a list and a counter of how many times this location occurs was started (see Figure 6).

[6] We would like to thank Gustavo Candela from the Biblioteca Virtual Miguel Cervantes for his assistance with this.


gustavo@gustavo: ~/Escritorio/sally-social
gustavo@gustavo: ~/Escritorio/sally-social

```
, 'Français': 1, 'CHULiege': 1, 'la France': 1, 'Le Soir Carte blanche': 1, 'Flandre': 1, 'Infectiologues': 1, 'LorientLeJour Mika': 1, 'de Lil
le': 1, 'Moderna': 1, 'Covid': 1, 'BouchesduRhône': 1, 'Corona': 1, 'France': 1, 'Ehpad': 1, 'Covid19\xa0  via lemondefr': 1, 'La Chine': 1})
gustavo@gustavo:~/Escritorio/sally-social$ python3 NER-French-analysis.py
Counter({'Belgique': 9, 'RT JdS_SK': 5, 'Houthulst': 2, 'Europe': 2, 'Bruxelles': 2, 'Japon': 2, 'EtatsUnis': 1, 'EpiCURA_hopital': 1, 'Ath': 1
, 'Français': 1, 'CHULiege': 1, 'la France': 1, 'Le Soir Carte blanche': 1, 'Flandre': 1, 'Infectiologues': 1, 'LorientLeJour Mika': 1, 'de Lil
le': 1, 'Moderna': 1, 'Covid': 1, 'BouchesduRhône': 1, 'Corona': 1, 'France': 1, 'Ehpad': 1, 'Covid19\xa0  via lemondefr': 1, 'La Chine': 1})
gustavo@gustavo:~/Escritorio/sally-social$ python3 NER-French-analysis.py
Counter({'Belgique': 36, 'EpiCURA_hopital': 10, 'Corona': 10, 'RT JdS_SK': 9, 'Ath': 8, 'Europe': 8, 'Covid_19': 7, 'France': 5, 'la France': 4
, 'Bruxelles': 4, 'Flandre': 3, 'Covid': 3, 'USI': 3, 'EtatsUnis': 2, 'Français': 2, 'Houthulst': 2, 'Japon': 2, 'Italie': 2, 'Soyons': 2, 'Hai
naut': 2, 'Namur': 2, 'CHULiege': 1, 'Le Soir Carte blanche': 1, 'Infectiologues': 1, 'LorientLeJour Mika': 1, 'de Lille': 1, 'Moderna': 1, 'Bo
uchesduRhône': 1, 'Ehpad': 1, 'Covid19\xa0  via lemondefr': 1, 'La Chine': 1, 'CHIRECgroup': 1, 'Angleterre': 1, 'Health healthcare': 1, 'Wuhan
': 1, 'JdS_SK': 1, 'ligne Covid': 1, 'lEurope': 1, 'Ribera': 1, 'Sicilia': 1, 'Liban': 1, 'Belles': 1, 'mifévrier\xa0  coronavirus': 1, 'Mol':
1, 'OrbanDoc Belles': 1, 'Londres': 1, 'ARNm Moderna': 1, 'Covid19': 1, 'Légère': 1, 'Allemagne': 1, 'La France': 1, 'Ittre': 1, 'BO': 1, 'Ath
 Rejoignez': 1, 'HealthBuzzBE': 1, 'ÉtatsUnis': 1, 'Québec': 1, 'RT LEXPRESS': 1, 'Celeval': 1, 'Covid19  en Calabre': 1, 'YaleSOM B_Hf_T': 1,
'LN24': 1, 'Le_Figaro': 1, 'Suède': 1, 'EpiCURA': 1, 'Ath ModernitéConfort': 1, 'Covid19 Il na': 1, 'DocWissam': 1, 'lodorat': 1, 'Etats': 1, '
Naples': 1, 'USAelection2020': 1, 'FranceDammel': 1, 'BW': 1, 'RT': 1, 'slatefr': 1, 'J90': 1, 'lEurope Covid_19': 1, 'confinement2': 1, 'Braba
nt Wallon': 1, 'Brabant du Hainaut': 1, 'Belgique COVID': 1, 'Dix Lunes': 1, 'Ath  IHAB': 1, 'Hôpitaux': 1, 'Nieuwsblad': 1, 'Libramont': 1, 'L
eSoumois': 1, 'pays de Galles': 1})
gustavo@gustavo:~/Escritorio/sally-social$
```

Figure 6: Output of the counter of locations detected in the French dataset of the BESOCIAL mini-pilot database

The steps described above were executed by a computational researcher who was not familiar with Twitter nor did he have any proficiency in French. These two limitations unearthed some important takeaways for heightening the potential reuse of the data.

First, more information should be provided on the rather cryptic table names (which are copied directly from the Twitter API) such as e.g. 'user_followers_count', ',user_listed_count', ' user_statuses_count', etc. This can be mitigated by providing a description for each of these in the respective README-files or by including a hyperlink to the Twitter documentation.

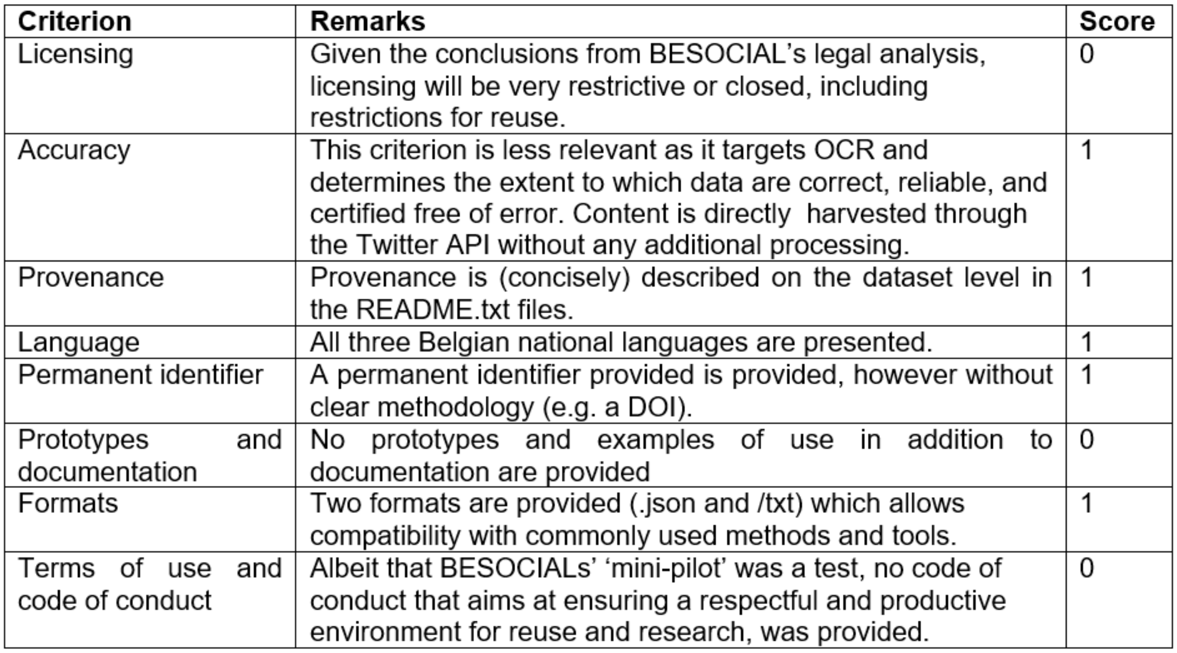
Second, as the researcher had no proficiency in French, more contextual information should be provided in advance on what the data set is about in order to identify the specific domain knowledge needed. In short, not knowing the context of the data and not having the required domain knowledge proved to be an important inhibitor in grasping and analysing the data.

3.3 Results from the quality assessment using Candela et al.'s (2021) benchmark

We tried to apply the benchmark developed by Candela et al. (2021) (see section 2.3) to the COVID19 related content that was harvested during BESOCIALs' 'mini-pilot' (and to how this content was presented as .json and .csv files). This exercise resulted in a total score of 5, see Table 1.

When comparing this results to those obtained by Candela et al. (2021) one needs to take into account some three important reservations, namely (i) the database used (with COVID19 related content that was harvested during BESOCIALs' 'mini-pilot') was a first test and was not made accessible in a 'production' or 'real live' environment, (ii) the benchmark developed by Candela et al. (2021) was initially create for a narrower scope namely to select datasets for computationally-driven research applied to Spanish text corpora, and (iii) the databases or libraries benchmarked by Candella et al. (2021), e.g. Chronicling America or Biblioteca Digital del Patrimonio Iberoamericano, target different domains and audiences and have completely different targets, infrastructures etc.

Nevertheless, plotting the benchmark of the BESOCIALs' mini-pilot database against the polar chart by Canella et al. (2021) that shows the highest and lowest benchmark scores respectively in their study, shows that the BESOCIALs' mini-pilot database scores quite satisfactory except on the criteria related to providing access ('license' and 'terms of use') and the criterion with regards to 'prototypes and documentation' (see Figure 7).



Criterion	Remarks	Score
Licensing	Given the conclusions from BESOCIAL's legal analysis, licensing will be very restrictive or closed, including restrictions for reuse.	0
Accuracy	This criterion is less relevant as it targets OCR and determines the extent to which data are correct, reliable, and certified free of error. Content is directly harvested through the Twitter API without any additional processing.	1
Provenance	Provenance is (concisely) described on the dataset level in the README.txt files.	1
Language	All three Belgian national languages are presented.	1
Permanent identifier	A permanent identifier provided is provided, however without clear methodology (e.g. a DOI).	1
Prototypes and documentation	No prototypes and examples of use in addition to documentation are provided	0
Formats	Two formats are provided (.json and /txt) which allows compatibility with commonly used methods and tools.	1
Terms of use and code of conduct	Albeit that BESOCIALs' 'mini-pilot' was a test, no code of conduct that aims at ensuring a respectful and productive environment for reuse and research, was provided.	0

Table 1: Benchmark of BESOCIAL mini-pilot database based on Candela et al. (2021)

To mitigate this, for the criterion 'license', close licenses which are less permissive and limit the usage should be avoided in favor of open licenses such as CC BY (Creative Commons Attribution License) or CC BY-SA (Creative Commons Attribution-Share Alike). To improve the criterion 'terms of use' (more) information on the terms of use of the BESOCIALs' mini-pilot database should be included as this is crucial to facilitate its reuse (Padilla et al., 2019)[7]. The last criterion which received a low benchmark score, 'prototypes and documentation' can be improved by simply providing examples of code or scripts that can be applied to the database in order to clean, search or analyse the data in the BESOCIALs' mini-pilot database. Also, examples

[7] Padilla, T., Allen, L., Frost, H., Potvin, S., Russey R., E., & V., Stewart. (2019). 50 Things --- Always Already Computational: Collections as Data. Zenodo. https://doi.org/10.5281/zenodo.3066237

of use, e.g. a published article that uses the BESOCIALs' mini-pilot database as a data resource, in addition to documentation can raise this benchmark score and facilitate the reuse of the dataset.

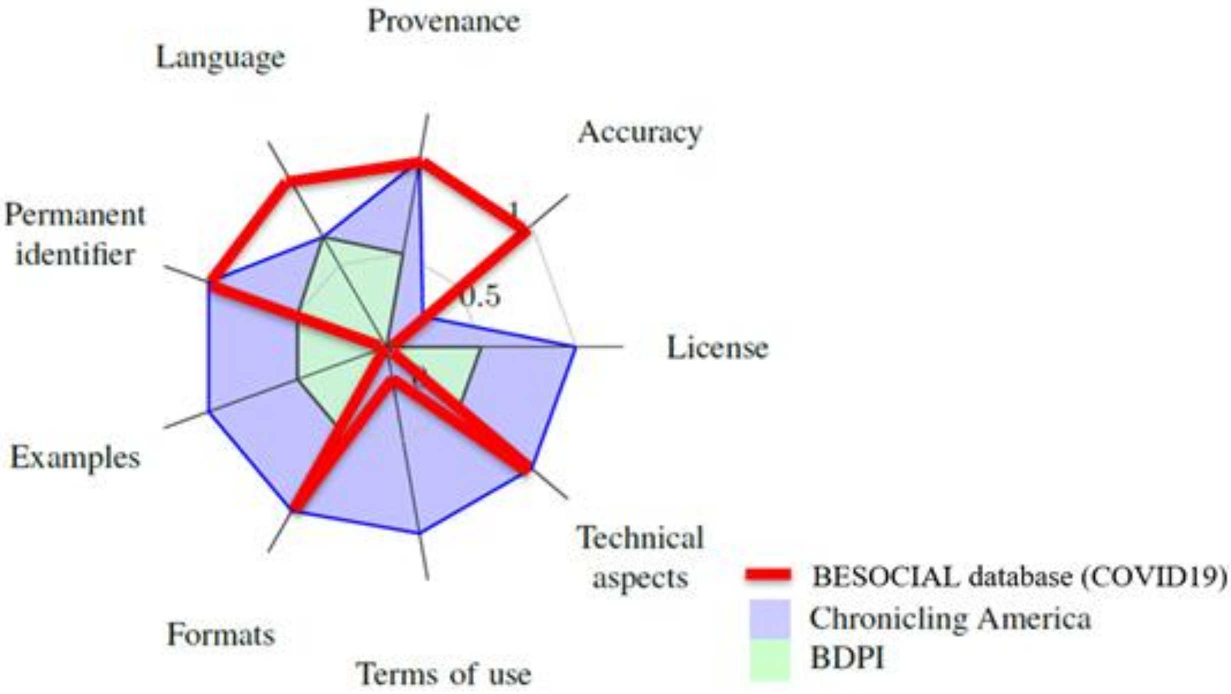


Figure 7: Polar chart based on Candella et al. (2021) with BESOCIAL mini-pilot database

4. Conclusion

This report focused on assessing the quality of the harvested content in BESOCIAL. In specific, we looked at the content that was harvested during BESOCIALs' 'mini-pilot' (Twitter COVID19 related content), and at how this content was opened up as 'data dump', i.e. a zipped folder containing various subfolders with the harvested data as .json and .csv, including the README.txt file.

In order to do this, we took a pragmatic three folded approach. A first line of inquiry into the 'quality' of the harvested content involved a qualitative research design that started from the different persona that were developed in WP 2 (see section 2.1). We created narratives – based on in-depth interviews – on how the persona would work with the provided harvested content and did a SWOT analysis technique using the perspectives of the persona's.

Next to this persona-driven research approach, we also took a more computational approach to assess the usability and quality of the content of the BESOCIALs' harvesting 'mini-pilot' by experimenting and exploring the harvested data in Tableau as well as Jupyter notebooks, two out-of-the-box or off-the-shelf software tools that are suitable for exploring and visualizing data (see section 2.2).

The third and last line of inquiry in our three folded approach involved using the benchmark of datasets for computationally-driven research developed by Candela et al. (2021)n (see section 2.3).

Our first persona driven approach showed some clear results. Most notably, it highlighted the practical and down-to-earth issues that would impede persona such as Manou and Jan (both with rather low computer and coding skills) in working with the harvested content and it underlined how straightforward and easy this job would be for a persona such as Febe (who has high computer and coding skills). In specific, an analysis from the perspective of Manou and Jan showed for example that the sheer size of the data limits their possibilities of exploring the data. This is also the case for the very cryptic and complicated filenames or the lack of clear and exhaustive documentation about the harvested corpus. While the lack of good documentation was also noted by the persona Febe, contrary to the other personas, the json-dump of the COVID19-collection harvested during BESOCIAL felt very 'natural' to her as she had the tools and expertise to process, clean and analyse such data. Our analysis shows that prior experience with .csv or .json-files, or more generally, data literacy is key and vital for managing, accessing and critically analyzing data and the data-collection process.

The Jupyter notebooks-case showed that people who are proficient in Python and in working with the Jupyter environment do not encounter many hurdles working with the harvested data. This case however also showed that, for researchers with no experience with Twitter or for researchers not proficient in the languages included in the database, more contextual information should be provided in advance on what the data set is about in order to identify the specific domain knowledge needed. In short, not knowing the context of the data and not having the required domain knowledge proved to be an important inhibitor in grasping and analysing the data. The Tableau-case showed similar results as it demonstrated that potential users with skills in working with Tableau can easily create basic visualisations of the harvested content.

In our third and last line of inquiry we tried to apply the benchmark developed by Candela et al. (2021) to the content harvested during BESOCIALs' 'mini-pilot' (and to how this content was presented as .json and .csv files). This exercise resulted in a total score of 5. It showed that the criterion 'license', (close licenses which are less permissive and limit the usage), the criterion 'terms of use' and the criterion 'prototypes and documentation' can be improved further (e.g. by providing examples of code of scripts that can be applied to the database in order to clean, search or analyse the data or by providing examples of use, e.g. an published article that uses the database as a data resource).