



WP4 Report - Evaluation of the Belgian pilot social media archive (Task 4.3)

Editors	Peter Mechant, Tim Theys & Eveline Vlassenroot
Responsible partners	MICT (UGent) GhentCDH
Version	Final
How to cite this?	P. Mechant, Theys, T. & E. Vlassenroot, BESOCIAL: Evaluation of the Belgian pilot social media archive (Task 4.3), June 2022.

Table of contents

1. Introduction.....	4
2. Feedback on the current interface.....	5
2.1 Simple search interface.....	6
2.1.1 <i>Orientating</i>	6
2.1.2 <i>Auditing</i>	7
2.1.3 <i>Constructing</i>	8
2.2 Advanced search interface	10
2.2.1 <i>Orientating</i>	10
2.2.2 <i>Auditing</i>	11
2.2.3 <i>Constructing</i>	12
2.3 Persona specific remarks	14
2.3.1 <i>Bart (Post-doc researcher)</i>	14
2.3.2 <i>Febe (PhD-student)</i>	14
2.3.3 <i>Ben (Journalist)</i>	15
2.3.4 <i>Manou (Scientific researcher)</i>	15
2.3.5 <i>Jan (Retired)</i>	15
3. Suggested improvements for the interface.....	15
4. Discussion and conclusion	19

Table of figures

Figure 1: Issues with the simple search interface.....	6
Figure 2: Unclear about (de)selecting a certain corpus	7
Figure 3: Search button indicating a search is being executed	8
Figure 4: Issues with results visualization interface.....	8
Figure 5: Issues with the advanced search interface	10
Figure 6: 'Unrealistic' filtering based on time	11
Figure 7: Search interface of TweetSets at https://tweetsets.library.gwu.edu/	11
Figure 8: The menu 'options avancées' in the advanced search interface	12
Figure 9: Issues related to generating data export	13
Figure 10: Mitigating issues 1, 3 and 7	16
Figure 11: Mitigating issue 2.....	17
Figure 12: Mitigating issue 5 and 10	17
Figure 13: Mitigating issue 9	18
Figure 14: Mitigating issue 5, 11, 12, 13 and 14	18

1. Introduction

This report describes the results of the BESOCIAL Task 4.3, titled 'Evaluation of the Belgian pilot social media archive'. The goal of this task is to evaluate, gain insights and provide feedback on the functionalities, look and feel and usability of the Belgian pilot social media archive-interface. These insights can then be taken into account when the interface to the social media archive is redesigned or further elaborated on.

After various meetings with the involved consortium partners in March 2022 it was jointly decided to tackle this task by means of an internal expert-review. This internal expert-review of the Belgian pilot social media archive-interface was conducted by 3 professionals in the field during 4 workshops in May and June 2022.


In order to structure the document, we first decided to focus on providing feedback on the various aspects of the current interface. A second chapter of this report then elaborates on how this current interface can be modified or adapted to even better serve the needs and requirements of potential users in terms of functionality and usability. The document ends with a short reflection and conclusion.

In order to structure the report further, we used - similar to T2.2 - the three conceptual devices (orientating, auditing and constructing) developed by Ogden and Maemura (2021). These 'devices' describe common research practices and associated challenges and highlight the significant time and energy required on the part of researchers to begin using archives. Rather than presenting a linear workflow or fixed set of practices, their concepts orientating, auditing and constructing necessarily overlap:

- 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. Hence, this phase includes getting to know the archive and its interface.
- 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. Hence, this phase includes creating a search query and filtering the dataset by focusing on a particular subset of the web archive; this can be accomplished by content, metadata, or some extracted information.

- 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. This phase thus includes extracting (after selecting a subset of material, the scholar typically then extracts some information of interest) as well as aggregating (the output (a collection of records of interest) needs to be aggregated or summarized).

2. Feedback on the current interface

In order to structure the evaluation results, we will first discuss the 'simple' search interface at  and will then elaborate on the insights gained by evaluating the 'advanced' search interface. To make things clear for the reader we decided to number the encountered issues and to visualise these on screenshot of the interface(s), see Figure 1, 4, 5 and 9.

2.1 Simple search interface

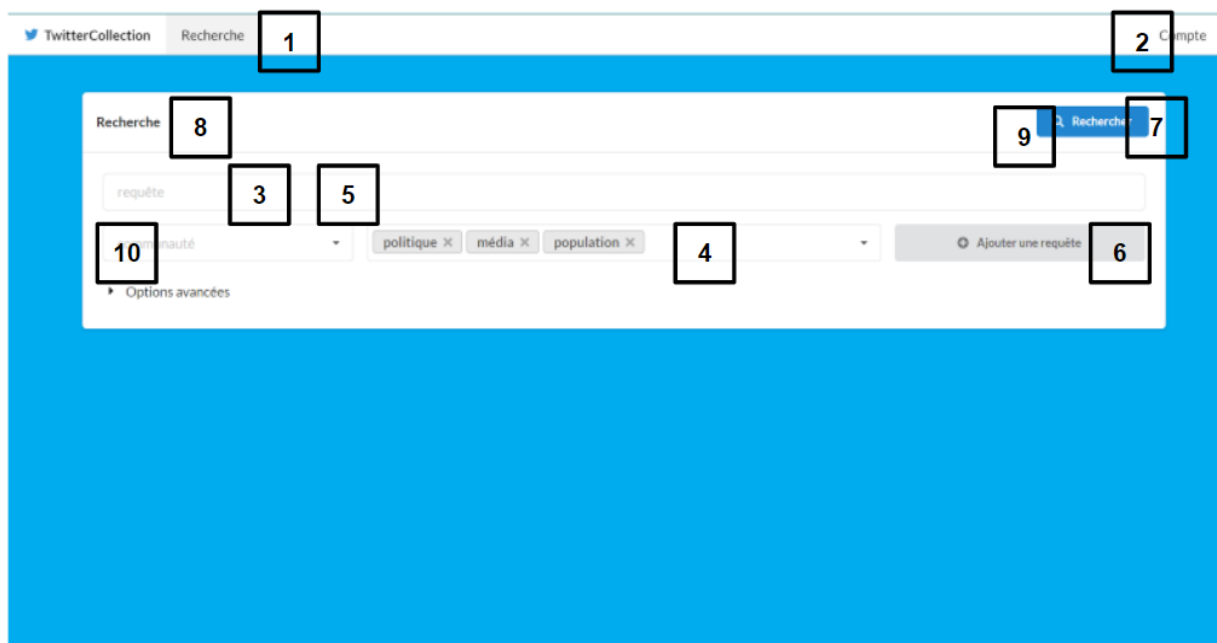


Figure 1: Issues with the simple search interface

2.1.1 Orientating

We used a broad interpretation of the dimension or phase 'orientating' and as such, it generally refers to getting to know the archive and its interface. We noticed some hurdles (issue 1 to 4, see Figure 1) that hinder the phase of 'orientating', basically because *too little contextual information* is provided to the interface user about e.g. the content of the archive or about how 'themes' were allocated to the archive collection.

1. A new user that visits the website and user interface without any prior knowledge needs contextual information on what he/she can do here, on what data is available through the user interface etc. Also provenance information about the archive collection should be provided such as information on how (and by whom) the tweets were collected (and when).
2. After login, a user should be able to easily access his/her profile information. It would be nice to expand upon the current functionalities and also include functions such as 'consulting saved queries', 'consulting search history', or a 'language toggle' to switch the user interface language).
3. Although input fields are described with a very concise help-phrase *in* the input boxes, these disappear when a user starts entering text. Thus input fields should be clearer

defined using a label above each input field.

4. A user is by default prompted to (de)select a certain corpus. It is not clear however how these themes were assigned - nor what they mean exactly - and what impact they have on the search query (see Figure 2).



Figure 2: Unclarity about (de)selecting a certain corpus

2.1.2 Auditing

Similar to our interpretation of Ogden and Maemura's (2021) first phase, we considered auditing as a broad process encompassing creating a search query and filtering a dataset. Most hurdles detected in this phase (issue 5 to 10, see Figure 1) are related to the *confusion* that a first user experiences due to the *different search and filter options*.

5. The user is not assisted in creating his or her search query. Options to mitigate this might include offering 'Auto Fill' functionality based on recent search queries or to include an affordance that enables users to save and name their search queries (these should then be visible in the user's profile information).
6. It is not clear how the functionality "Ajouter requete" differs from adding two search phrases in the input field and this causes confusion.
7. A typical end user expects the button "Search" or "Execute" or in this case, "Rechercher" to be positioned below the input fields and not above them.
8. It is not clear for the user which input fields are mandatory and which are not.
9. Although the interface visualizes when a search query is being executed (see Figure 3) by changing the appearance of the search button, this is not visually very clear or prominent. Also, an option to cancel the search query if it takes too long should be included here.
10. The final interface element that causes confusion is the language filter. On the one hand it is not clear what this filter actually does (e.g. does it filter tweets from Belgians or tweets geolocated in Belgium or does it filter tweets based on language?). It is also not clear why 'Belgique (nl)' is included twice nor why 'all' is not included or even selected

as default.

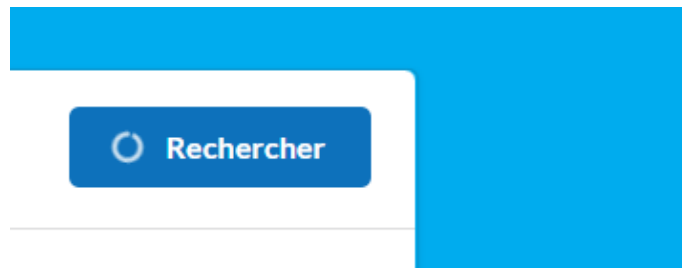


Figure 3: Search button indicating a search is being executed

2.1.3 Constructing

In line with our previous interpretations of Ogden and Maemura's (2021) vocabulary, we consider 'constructing' broadly as encompassing consulting search results and activities surrounding the creation of a subset of data. Hurdles during these activities (issue 11 to 15, see Figure 4) mainly point to providing *too little details and interactive features*.



Figure 4: Issues with results visualization interface

11. While some information about the search result is provided, more useful information could be added here in a concise textual manner such as the date of the first and last tweet in the collection, the percentage of tweet types (e.g. retweet, reply, ...)... Also the number of search results seems to be limited to 10000 tweets, something the user is not informed about explicitly.
12. While the word cloud provides a clear overview of the search results, interactivity is missing; a user should be informed about the frequency and number of occurrences of the words when hovering above a certain word. Alternatively, clicking on a certain word should result in an overview of the corresponding tweets. No export-option to download and store the word cloud visualization is offered.
13. While the bar chart and pie chart visualizations certainly provide an added value to quickly grasp the search query results as a whole, they are specifically tailored towards this collection and might not be relevant for other corpora or for other researchers from different disciplines. Bar chart and pie chart visualizations can only be exported as .csv-files and not as an image.
14. Ideally users should immediately see a small sample of tweets that represent the search results rather than first having to click on a separate button. When this sample of tweets is then shown, their representation should try to approximate the original tweet. In line with this, embedded media, hyperlinks, user name (looks clickable but isn't) and the tweets themselves should be hyperlinked.
15. The export procedure seems on the one hand overly complicated; e.g. the user needs to enter his/her email when he or she is already logged in, or the user needs to 'request access' as the URL that directs him/her to the search results in a Google Sheets file only enables 'View Only'-mode. On the other hand, not enough options are provided to the user; e.g. ideally the user should be able to choose the export-format (.csv, .xlsx, .json) as well as the data fields that should be included in the export. In this regard is the current export rather limited (including only metadata fields such as id, date, segment, party, user_name, user_screen_name, retweet, matches, full_text, hashtags, mentions, urls and medias) while the Twitter API offers much more information for each tweet (e.g. the 'referenced_tweets'), see also <https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/tweet>.

2.2 Advanced search interface

While many of the issues mentioned in section 2.1 also apply to the more advanced search interface discussed below, and while - certainly for the constructing phase - this division into 'simple' and 'advanced' feels rather superficial, it serves its purpose by highlighting issues that create hurdles in the orientating, auditing and constructing phases of more 'advanced' users or users with more experience in working with archives. Similar to section 2.1 we have numbered the potential hurdles one can encounter during these activities (issue 1 to 8, see Figure 5).

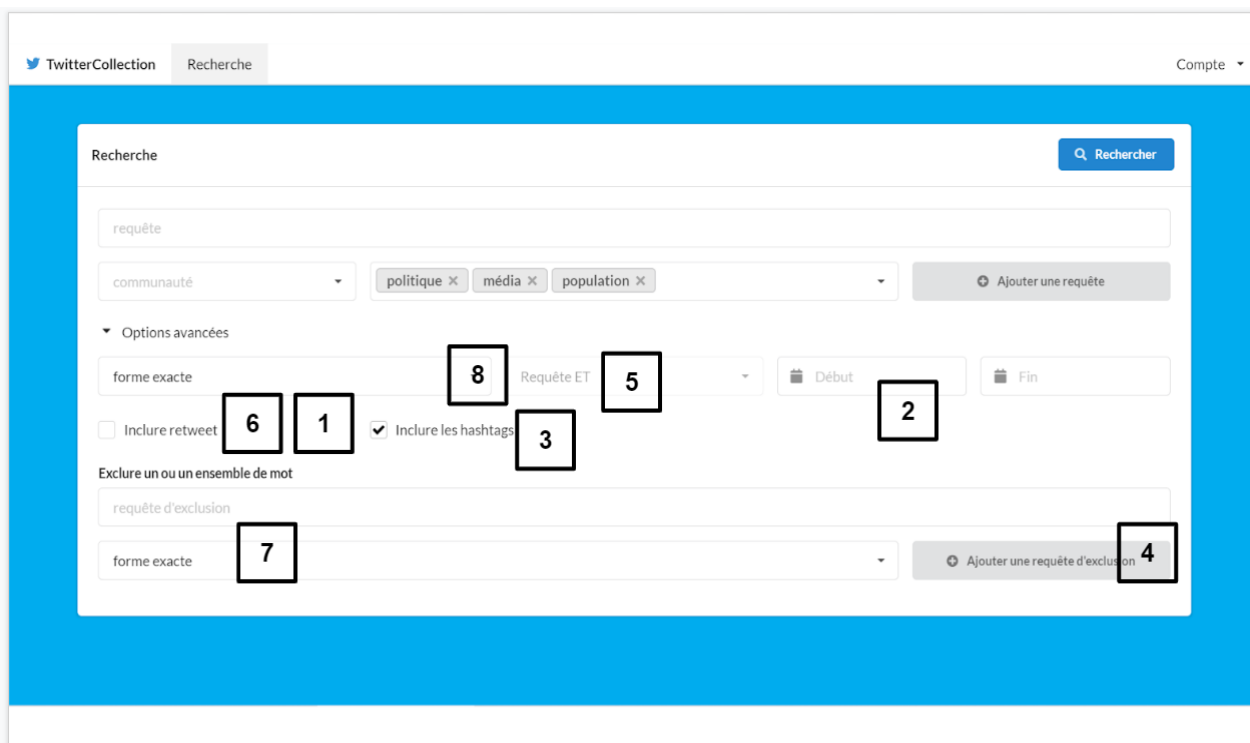


Figure 5: Issues with the advanced search interface

2.2.1 Orientating

Digging deeper into 'orientating' (getting to know the archive and its interface) two additional issues were noted;

1. Some of the interface options are confusing, e.g. it is not clear to what extent the advanced options (e.g. 'Inclure les hashtags') apply to only the first query or also queries that are added with the 'ET' or 'OU' option.
2. While the affordance to allow a user to filter the result set using time delineation is certainly useful we noticed that this functionality is not aligned with the actual content that is filtered on (e.g. a user can create a filter on '1891', see Figure 6). Also, providing

a user with some presets (e.g. last week, month, year) might be useful.

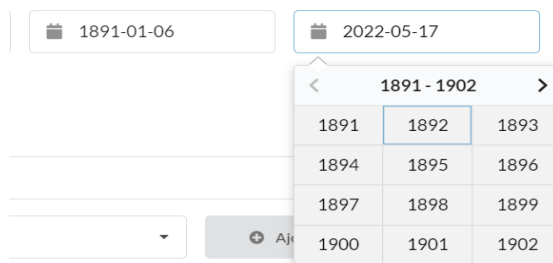


Figure 6: 'Unrealistic' filtering based on time

2.2.2 Auditing

Again, we consider auditing as a broad process encompassing creating a search query and filtering a dataset. Most hurdles detected in this phase (issue 3 to 7, see Figure 5) are related to the rather *confusing set-up with different fields* where 'hashtags' can be selected and the implementation of boolean operators ('ou' and 'et') in the search interface.

3. As already mentioned when discussing the simple interface, end-users find the different input fields where hashtags can be (un)selected rather confusing.

While the interface offers the option to only search on hashtags via the 'options avancées' menu, a more straightforward way of presenting the end users with these affordances would be to split the search fields in different sections while at the same time offering the option to exclude certain search phrases, similar to the interface of TweetSets (a service of George Washington Libraries (see Figure 7). In this way the button 'Ajouter une requête' can be omitted.

A screenshot of the TweetSets search interface. It features several search filter sections, each with a text input field and a small icon on the right. The sections are: 'Tweet text' with a sub-note 'For a retweet or quote, tweet text includes the text of the source tweet.'; 'Contains all' with a sub-note 'Text must contain all of these terms (AND). Comma separate multiple terms. Space separated words are treated as a phrase.'; 'Contains any' with a sub-note 'Text must contain one of these terms (OR). Comma separate multiple terms. Space separated words are treated as a phrase.'; 'Excludes' with a sub-note 'Text may not contain (NOT). Comma separate multiple terms. Space separated words are treated as a phrase.'; 'Hashtags' with a sub-note 'Text must contain one of these hashtags (OR). # is optional. Case insensitive.'; and 'Mentions' with a sub-note 'Any of these screen names must be mentioned (OR). @ is optional. Case sensitive.'

Figure 7: Search interface of TweetSets at <https://tweetsets.library.gwu.edu/>

5. In line with this previous remark, omitting the Boolean operators and presenting the end-user with input fields, titled 'Contains any', 'Excludes', etc., instead, makes the interface more transparent for the user.
6. The search interface does not exploit all filtering options that are possible on the archived collection, e.g. options to search on 'Posted by', 'Mentions', 'URL' or 'In reply to' are missing.
7. In a similar vein, the advanced search interface could offer more filter-options including filtering on tweets that 'Contains any URL', 'Is geotagged', 'Has at least one media (embedded images)'. More thorough filtering on Tweet types could also be enabled.

2.2.3 Constructing

We consider 'constructing' broadly as encompassing consulting search results and activities surrounding the creation of a subset of data. Here, the main issues detected are mainly related to *how the export is constructed and what information it provides*.

8. While the options under the menu 'options avancées' can be useful for certain users or researchers, in specific for those focusing on grammatical or language-related aspects, e.g. NLP researchers, offering these choices by default does not make much sense (see Figure 8).

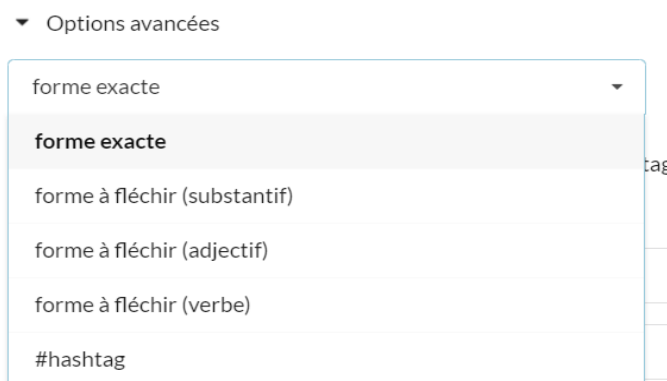


Figure 8: The menu 'options avancées' in the advanced search interface

9. As already mentioned previously, it would be beneficial if the end-user was able to save and name certain queries he or she executed and, in a similar vein, has the possibility to name the export he or she created (see Figure 9).

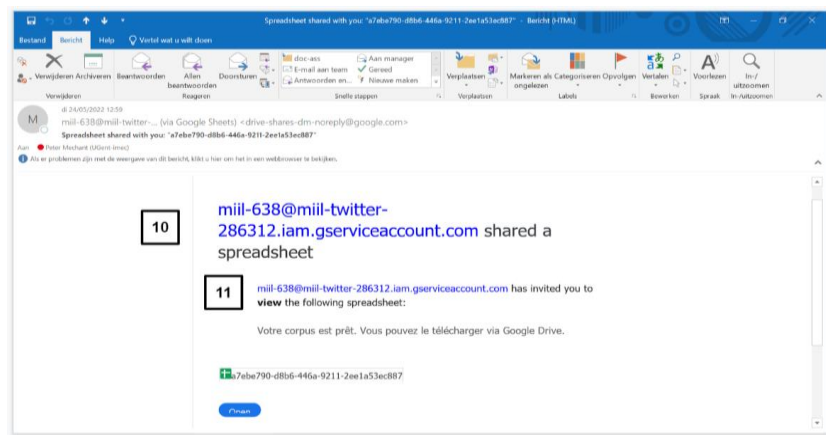
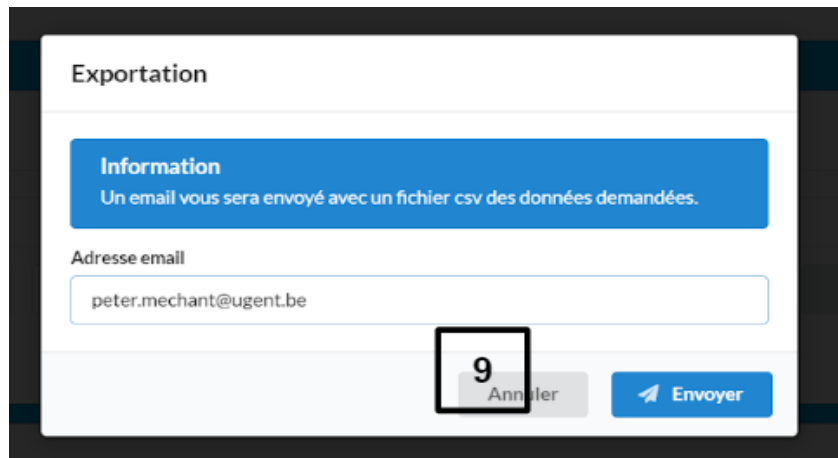


Figure 9: Issues related to generating data export

10. After exporting the data set the user receives a rather cryptic email on the address he/she had to input in the pop-up window. This email can be improved in terms of readability and clarity by making the subject of the email more explicit and by adding a text that explains how the Google Sheet was generated and what it contains.
11. Next to the readability of the email that notifies the user the export file is available on Google Sheets, accessing this export file could be improved by setting the rights by default on 'Edit' rather than on 'View'. The export should also include a 'readme'-file that describes and provides more information about the different columns in the export file. Finally, we noticed that all hashtags are aggregated in one column (while creating separate columns for each hashtag would be better) and that character formatting in the export is not always correct.

2.3 Persona specific remarks

Based on previously developed persona in the field of (web) archiving (see T2.2 and T3.2), and taking group discussions and the results of the PROMISE project into account, five persona were created earlier in the project (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

Below we briefly reflect, for each persona, on to what extent the current search interface and results visualizations (including export) meet the requirements of these persona, taking into account the research questions that they want to tackle.

2.3.1 Bart (Post-doc researcher)

While Bart has a need for advanced search options, the presentation of the results needs to be easily interpretable and exportable in a commonly used format. Given the fact that he “lacks the time to thoroughly learn/acquire new skills that are required” a self-explaining interface is needed. Ideally, he receives ready to use exports in an easily readable format (e.g., google sheets, .xlsx) or even preconstructed analysis that can be exported (including tables and charts) as Bart “has little to no expertise in working with big data nor expertise in working with analytical software that can process those big datasets”. For Bart’s specific research question he needs to be able to filter on tweets of specific accounts and he should be able to set time boundaries (“over the last year”) as a filter.

2.3.2 Febe (PhD-student)

Febe needs advanced search options and the export of raw data in a format that allows further advanced data analysis. She also needs the option to filter on language used in a tweet as she “needs to analyse and develop the methodology for several languages.”

2.3.3 Ben (Journalist)

Ben needs advanced search options but the results have to be easily interpretable and exportable in a commonly used format. Given the fact that “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.”, ready to use exports in an easily readable format or preconstructed analysis that can be exported (including tables and charts), is needed for this persona.

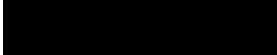
2.3.4 Manou (Scientific researcher)

Manou needs to be able to export huge data sets into a processable format. As such, she should be provided with the possibility to collect as much data as possible, preferably unlimited (this is now limited to maximum 10.000 tweets).

2.3.5 Jan (Retired)

“Jan has no programming expertise, nor is he acquainted with big datasets or analytical software”, thus he needs a rather simple search function and ready to use visualizations of the data. However, currently tables and charts are not exportable as images (now only low resolution screenshots are a way to overcome this).

3. Suggested improvements for the interface

While chapter 2 focused on listing the evaluation results of the 'simple' search interface as well as the insights gained by evaluating the 'advanced' search interface at  the current chapter will focus on how this current interface can be modified or adapted to even better serve the needs and requirements of potential users in terms of functionality and usability. In order to make this clear for the reader we reused the numbering of the encountered issues (see chapter 2), visualizing these potential improvements by means of screenshots of potential interface(s).

- The first screenshot (Figure 10) mitigates issues in the orientating phase (issue 1 & 3), in specific the presentation to the new visitor of contextual information on what he/she can do here, on what data is available through the user interface etc. (issue 1), clearly defining the input fields (issue 3), as well as issues during the auditing phase (the position of the 'search' button, issue 7).

- The second screenshot (Figure 11) shows how issue 2 might be tackled by providing more interactivity and information behind a user's profile information, expanding upon the current functionalities by including saved searches and exports.
- The third screenshot (Figure 12) addresses issues related to auditing or the process of creating a search query and filtering a dataset, in specific issues 5 and 10, by providing the user with more clarity or assistance when formulating the search query.
- The next screenshot (Figure 13) mitigates issue 9, in specific it makes clearer when a search query is being executed and allows to cancel the search query if it takes too long.
- The final screenshot (Figure 14) shows how several issues during the 'constructing' phase (activities surrounding the creation of a subset of data) and the visualization of the search results, in specific issues 5, 11, 12, 13 and 14, can be resolved.

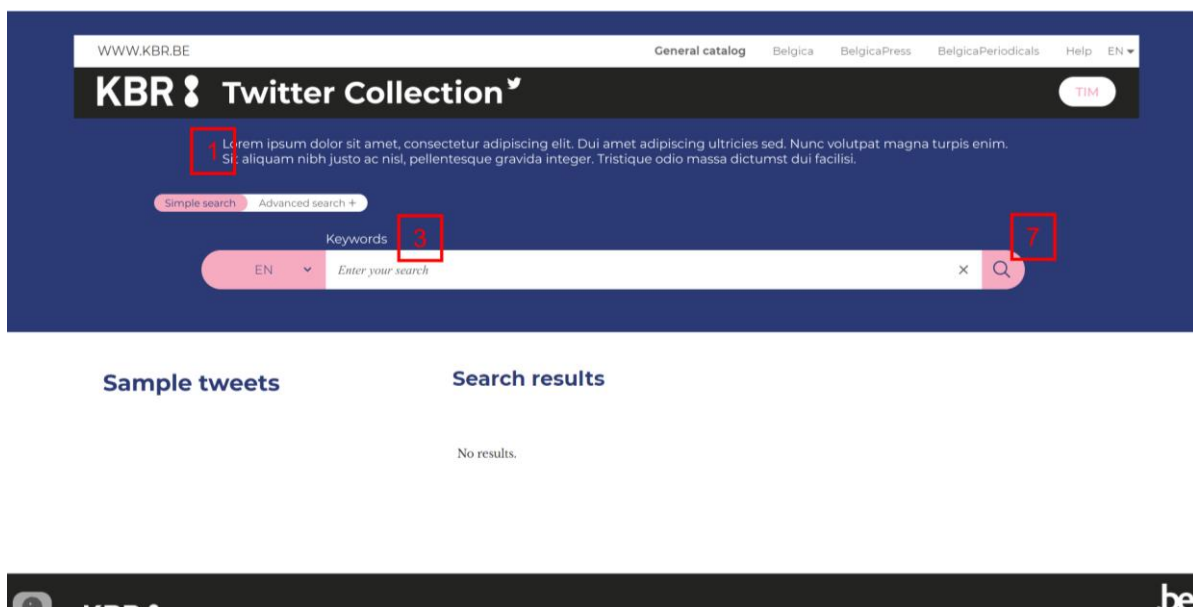


Figure 10: Mitigating issues 1, 3 and 7

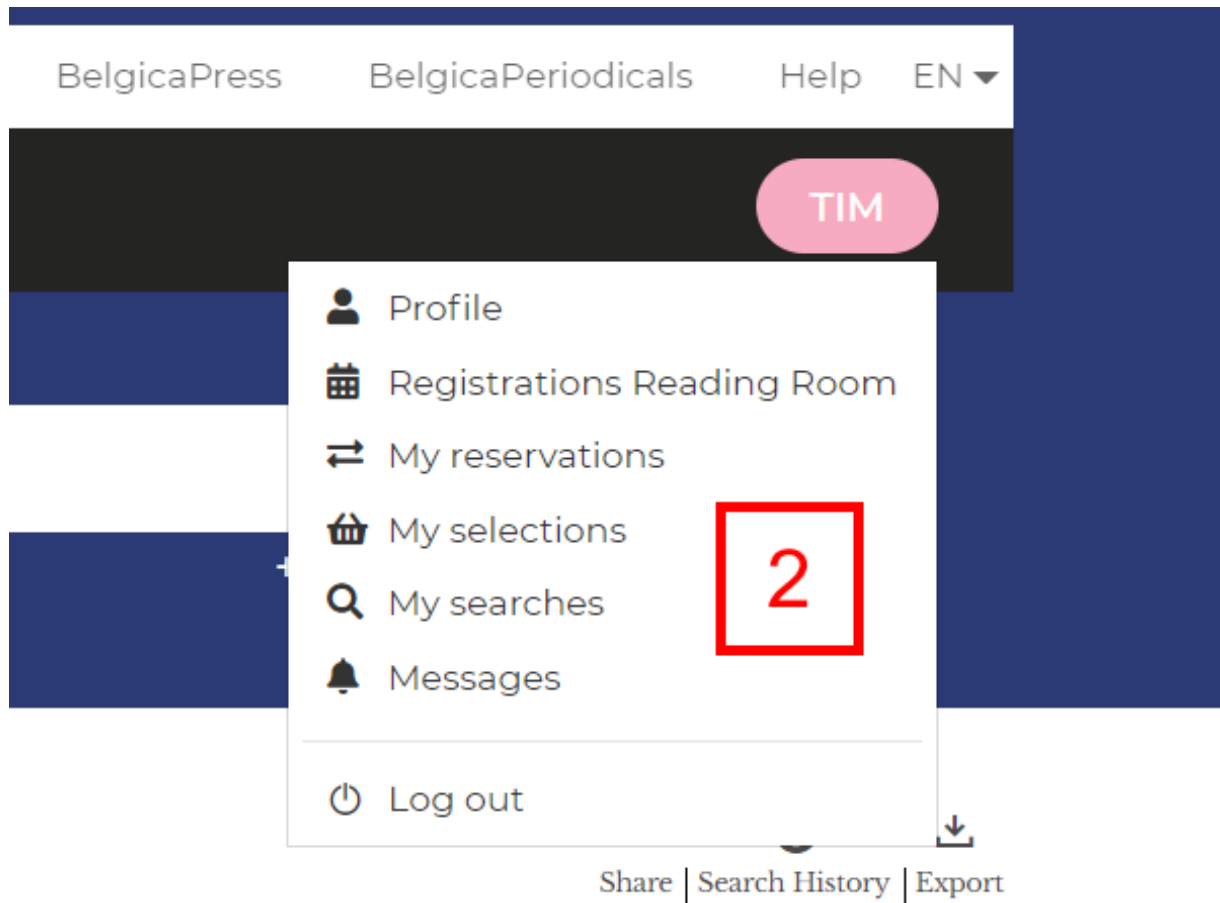
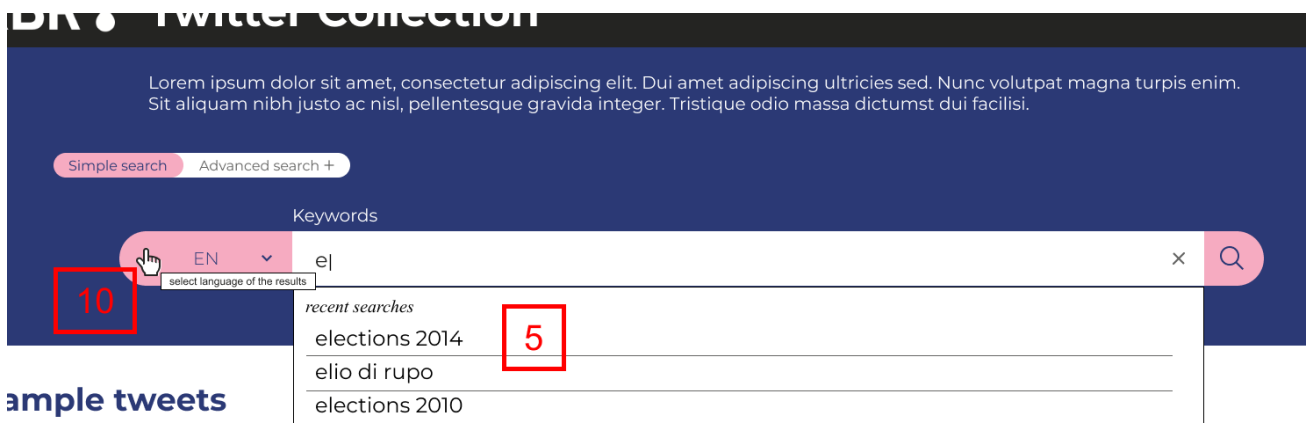


Figure 11: Mitigating issue 2



No results.

Figure 12: Mitigating issue 5 and 10

Sample tweets

Search results

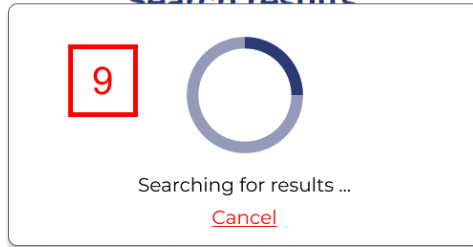
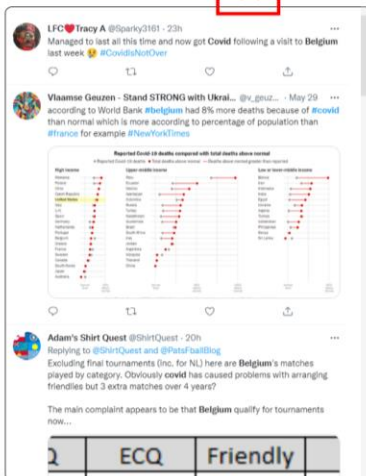


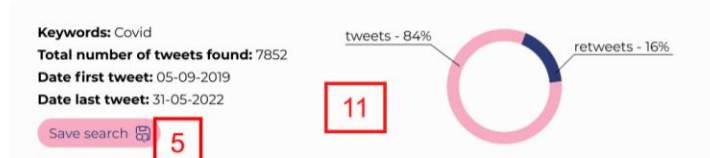
Figure 13: Mitigating issue 9

Sample tweets

14



Search summary



Wordcloud



Charts

Single word occurrences



Hashtag occurrences



Figure 14: Mitigating issue 5, 11, 12, 13 and 14

4. Discussion and conclusion

In this document we provided feedback on various aspects of the current BESOCIAL user interface, developed by CENTAL and available at [REDACTED]. This was done based on internal expert-reviews by 3 professionals in the field during 4 workshops in May and June 2022. In order to structure the results the simple and advanced search interface were discussed separately and the three conceptual devices (orientating, auditing and constructing) developed by Ogden and Maemura (2021) were used as guidelines. We also looked at the current interface from the perspective of five persona that were developed earlier in the BESOCIAL project.

While we acknowledge that the interface was initially developed in the context of NLP-research and for a specific target group (i.e. CENTAL researchers) and while we are aware that the current interface offers quite some functionality - and that these functions are directly related to the time and effort that can be spend in developing this front-end interface - we also included in the report suggestions on how this interface can be modified or adapted to even better serve the needs and requirements of potential users in terms of functionality and usability.

Main points-of-pain that were detected during the analysis include the fact that too little contextual information is provided to the interface user about the content of the archive or e.g. about how 'themes' were allocated to the archive collection. Other hurdles are related to the confusion that a first user experiences due to the different search and filter options, lack of interactivity and export-options (how the export is constructed and what information it provides).

Based on these detected points-of-pain (which were numbered in the document), chapter 3 of this report suggests some improvements and additional functionality that can be added to the BESOCIAL user interface, these include amongst others; adapting the search fields in order to avoid confusion by the end-user, adding more contextual information to the interface, offering the end-user more functionalities (e.g. to save a search query) and expanding upon and reworking the export functionality.