# BEHIND THE SCENES OF THE BELGIAN WEB ARCHIVE RESEARCH OPPORTUNITIES AND CHALLENGES

**Patricia BLANCO**
Internship at the Royal Library of Belgium during the PROMISE project
MSc in Digital Humanities at KU Leuven

■ When the *PROMISE* project was launched in Belgium to set up a national web archive, a researcher in Digital Humanities was allowed to participate in the different stages of the web archiving workflow: selection and harvesting of websites, quality control of the captures and access to the archived files to test research tools. The goal was to know the potential researchers' needs and expectations, but also the technical requirements and limitations of providing data-level access to the files. This article summarizes the best practices and the challenges that were identified during this researcher's experience.

■ Lorsque le projet *PROMISE* a été lancé en Belgique pour mettre en place une archive web nationale, un chercheur en *digital humanities* a été autorisé à participer aux différentes étapes du processus d'archivage web : sélection et récolte des sites web, contrôle de la qualité des captures et accès aux fichiers archivés pour tester les outils de recherche. L'objectif était de connaître les besoins et les attentes des chercheurs potentiels, mais aussi les exigences et les limites techniques pour fournir un accès aux fichiers. Cet article résume les meilleures pratiques et les défis qui ont été identifiés au cours de l'expérience de ce chercheur.

■ Bij de lancering van het PROMISE-project in België om een nationaal internetarchief op te zetten, mocht een onderzoeker in Digital Humanities deelnemen aan de verschillende fasen van de workflow voor internetarchivering: selecteren en verzamelen van websites, kwaliteitscontroles van die sites en toegang tot de gearchiveerde bestanden voor het testen van onderzoekstools. Het doel hiervan was om de potentiële behoeften en verwachtingen van onderzoekers te kennen, maar ook de technische vereisten en beperkingen van het verstrekken van toegang tot de bestanden op gegevensniveau. Dit artikel geeft een overzicht van de beste praktijken en van de uitdagingen die tijdens de ervaring van de onderzoeker naar boven kwamen.

## Introduction

The first initiatives to preserve websites were launched in the mid-90s by nonprofit organizations and national libraries. They were aware that the information published online and its form were unique, but also prone to disappear without a trace. The main motivation for national libraries was to preserve it, both for the general public and researchers. However, many people today are still unaware of the existence of web archives and their use has not yet been consolidated among the research community. Capturing, accessing and analyzing archived web data is still plagued by unknowns and challenges, both technical and legal.

The *PROMISE* project[1] was initiated by the States Archives and the Royal Library of Belgium (KBR) in 2016 to build a national web archive. Raising awareness and promoting its use for research became one of their priorities. A Master' student was invited to participate in the project and offer some feedback on the potential users' needs. The student created a collection of websites around a specific topic in order to explore different selection methods; assisted during quality control tests to identify the problems with the web harvesting tools and ultimately, explored the use of computational tools (text and hyperlinks extraction and analysis) to understand the technical requirements of giving data-level access to the archived files.

This article aims to shed some light on web archives and show what happens behind the scenes of a national web archive.

## Selection of web content

In countries with legal deposit laws, national archives are authorized to archive websites with patrimonial purposes without having to ask the author's permission. This allows them to launch large-scale crawls and systematically archive, for instance, every website with the country-code top-level domain (ccTLD): websites ending in ".be", ".brussels", ".vlaanderen", ".gent" for Belgium, ".es" for Spain, etc. However, there are other generic top-level domains (gTLD), such as ".net, ".edu", ".gov" or ".org", that might also contain information related to a country and its inhabitants.

For the creation of special collections, the selection of websites needs to be done manually. Under the topic "representation of minorities in Belgium", three collections were created in order to save websites related to the Spanish, the Italian and the Portuguese communities in the country.

Search engines were the main tool used to find relevant websites, with a special attention to those disseminating news, literature, cultural associations and events concerning these communities. The results were obtained through the combination of keywords that were translated into the official and co-official languages of Belgium, Spain, Portugal and Italy[2.]

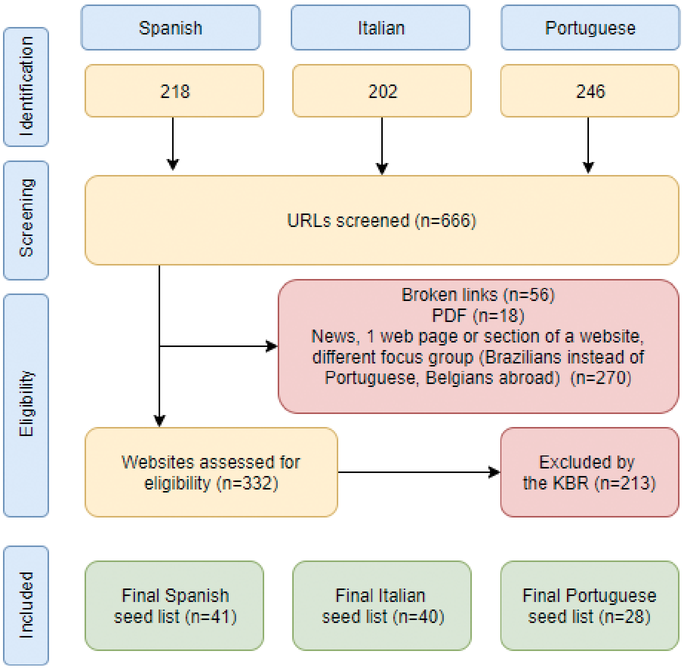
Identification
Screening
Eligibility
Included
Spanish
Italian
Portuguese
218
202
246
URLs screened (n=666)
Broken links (n=56)
PDF (n=18)
News, 1 web page or section of a website, different focus group (Brazilians instead of Portuguese, Belgians abroad) (n=270)
Websites assessed for eligibility (n=332)
Excluded by the KBR (n=213)
Final Spanish seed list (n=41)
Final Italian seed list (n=40)
Final Portuguese seed list (n=28)

*Fig. 1: Flow diagram illustrating the selection process of thematic collections.*

Social media platforms were excluded from these collections. They were, however, used as a crowdsourcing tool to retrieve relevant websites. A call for participation was made in different Facebook groups of these communities: users were asked to share the websites they use to connect with their compatriots in Belgium, share experiences, promote their language and culture, organize events, etc. By encouraging participation, we also brought attention about the existence of our web archive.

From the initial 666 URLs screened, only 109 were selected for the crawl (fig. 1): some websites were no longer accessible and those excluded did not fit into the KBR's cultural heritage preservation mission (small businesses websites) or were not sufficiently relevant to be archived entirety (one-page news articles).

## Research opportunities

These collections are not only a testimony of the presence, history, impact and evolution of foreign communities in Belgium, but a corpus from which other countries can benefit to carry out transnational research.

The largest existing text corpus is the Web. It is a gold mine for the field of computational linguistics. These collections contain text in ten different languages[3] and multiple dialectal variations. The text can be automatically indexed to allow full-text search and text analysis. Since it is machine readable, we can take advantage of Natural Language Processing (NLP) tools to extract data (Named Entity Recognition), and improve natural-language understanding and natural-language generation tools (machine translation). This multilingual corpus also allows the study of language varieties, such as dialects, sociolects and multiethnolects.

Although commercial pages have not been included, directories of companies and organizations have been kept. They contain addresses of small businesses and professionals, trade unions and cultural associations created by emigrants in Belgium.

With this information we can map their location and observe the geographical distribution of these communities across the country, explore their sectors of activity and study their socio-economic evolution.

## Harvest and Quality control

There are a number of open-source tools we can use to harvest websites. However if we want to work with a large sample of websites, we still rely on institutions with the capacity to regularly archive and preserve millions of them. Some organizations, such as the Internet Archive or Common Crawl, opt for a discovery crawl: they feed the web crawler a list of websites to be harvested, let it follow the hyperlinks that they contain, add them to the list and continue to increase the scope. Their goal is to automatically discover new websites and preserve the maximum possible of the web sphere.Cultural institutions tend to limit the scope to the websites of their selection, particularly for thematic collections. Used by most

| | Spanish | Italian | Portuguese |
|---|---|---|---|
| Initial seed list | 41 | 40 | 28 |
| URLs downloaded | 1,772,554 | 2,610,369 | 2,604,372 |
| URLs queued | 87,438 | 7,938,266 | 14,166,405 |
| Size | 81 GB | 163 GB | 865GB |

*Tab. 1: Harvest progress: initial seed lists, URLs downloaded and queued, and size of the collections.*

national libraries, Heritrix was also the web crawler used by the *PROMISE* project to capture extracts of the Belgian web. Heritrix visits all the hyperlinks in our list ("seed list"), capture their content and save it into multiple container files, with a maximum size of 1 GB each. These container files are known as WARC files (Web ARChive), a standard file format for web archival.

## Problems and possible solutions

We identified the first problems in the harvest when the size of the Portuguese collection, initially with less URLs, was increasing faster and slowing the download (Tab. 1). When we analyzed the WARC files, we noticed that many contents were videos from a Portuguese news channel in Belgium. This made us re-evaluate our selection strategy and the parameters set on web crawler. These can be configured to restrict the capture of certain files, such as audio and video, but that would also mean that our archived versions would be incomplete.

Can we calculate the size of a crawl in advance? As far as we know, there are no specific tools available. One of the solutions could be launching a crawl to collect only the metadata of the websites in our list, without archiving the contents. The metadata about the size of the files would provide an estimation of the space we will need. We could then decide if we restrict the capture of some files on certain websites or if we excluded them from the seed list.

We replay the websites to see if their appearance matches the live version and if all the contents had been archived correctly. We identified the following issues in a number of websites: variations in the fonts, the page layout had not been captured and the contents had been displaced, images or icons were replaced by symbols or disappeared, making some drop-down menus undiscoverable, etc.

The root of these problems stems from the limitations of the software itself, especially when capturing inline JavaScript and dynamic content. It also depends largely on the "archivability" of the website which determines how easy it is to archive. If standards are not respected in the design of a website, this will not be 100% archivable. There are tools to evaluate the archivability of a website[4], a practice that should be generalized among web builders and web archivists. Websites poorly designed slow down and even block the web crawlers. Others websites could not be captured due to the robots.txt file. This is a file, added by the author or the web designer, to prevent web crawlers from harvesting the entire website or sections of it.

## Using web archives

The most straightforward way to access archived websites is by interacting with replayed versions of the web pages[5]. With some limitations, we can navigate the old version as we would do it with a browser. We can observe their appearance, read its contents, but we cannot process the data or do quantitative analysis.

For any organizations, the challenges of giving access to the web data are not only legal but also technical. The legal deposit serves to justify archiving websites for patrimonial purposes, but other legal restrictions play a part concerning the publication of this data[6].

There are tools that allow us to concatenate WARC files and extract only the data we need, such as text, images, metadata and hyperlinks. A sample of the three collections (Tab. 2) was placed on a remote server and we used the *Archives Unleashed Toolkit* (AUT), a command line-based tool, to explore it. One of its features is to identify the types of files and file formats captured. Thousands of files could not be recognized among the most common or standardized formats, which reveals the problems with long-term preservation (LTP) once these files have to be migrated.

We also used the AUT to extract hyperlinks and text filtered by language or by web domain.

| | Spanish | Italian | Portuguese |
|---|---|---|---|
| Number of WARC files | 22 | 11 | 36 |
| Total size | ~ 70 GB | | |

*Tab. 2: Number of WARC files and size of the sample used to explore the collection.*

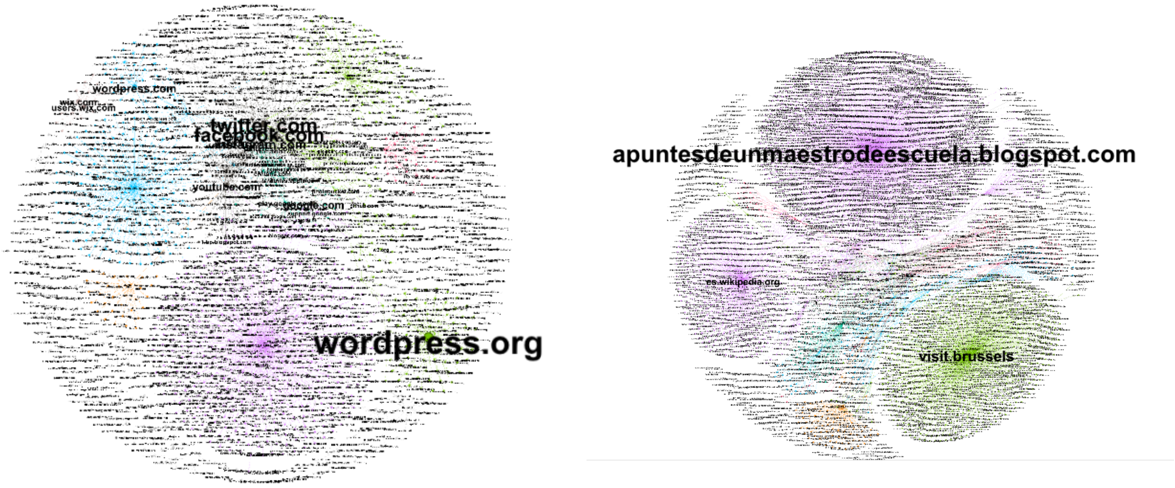

*Fig. 2: Data visualizations to explore the hyperlink networks in our collections.*

Hyperlinks can be visualized with tools such as *Gephi*[7]. We could observe which domains have more presence in our collection but also how they link to each other and towards other websites. The study of the hyperlinks can reveal strong connections between websites that we would not be able to appreciate with close reading methods.

For example, some researchers have studied hyperlinks to reveal connections between online newspapers and political parties[8].

The websites more frequently linked to from our collection were Wordpress pages and social media platforms, such as Facebook, Twitter and Youtube. Some of the websites we selected have links to their social media profile or Facebook group. These are widely used by the expats and are updated more frequently than the website, which sometimes ends up being completely replaced. We cannot overlook the importance of social networks and the need to archive some of their publicly accessible content too.

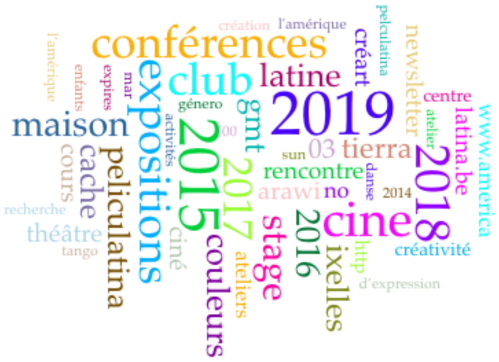

*Fig. 3: Visualizing text from the website La Maison de l'Amérique Latine*[9].

Extracting text was more challenging and the 36 WARC files from the Portuguese collection often caused the AUT software to close. The output text by language requires cleansing in order to be analyzed. However, we used text extraction tools to explore the most frequent words from each domain (fig 3). Some can be included in the description of the website in our web archive catalog as keywords or tags.

## Conclusion

The development of a big data infrastructure proved to be necessary to let people download, manipulate, analyze or extract GBs of data to carry out their research. However, the priority today is to save the websites. We should first ensure their preservation, so that they can be used once the legislation and technology ease working with this re-born digital source.

And this is a task to which we can all contribute:

- We should use archived versions of websites more often in our references. Archived versions are like editions of a book. The archive date is even stamped on the hyperlink. We can find them in web archives or even create them ourselves with a simple click[10].
- Any organizations, large or small, should take advantage of open source tools to create their own web archive. This will allow more diversity in the sources available in the future.
- We must respect standards when designing websites to make our pages "archivable" and restrict the use of robots.txt files when it is possible. We can send our websites to web archives to be preserved, by sending the hyperlink or the files. Institutions can consider embargoes to give

access only when the author authorizes it, or when the page is no longer accessible online.

- National libraries can offer small samples and curated collections, tutorials to introduce researchers to the web archives and promote their use.

Saving websites asks for a collaborative effort. We can all help to save relevant information and contribute to create sustainable web archives.

**Patricia Blanco**
*KULeuven*
Oude Markt 13 -3000 LEUVEN
p.blanco.nunez@gmail.com

April 2020

## References

ACKLAND, Robert; GIBSON. Hyperlinks and networked communication: a comparative study of political parties online. *International Journal of Social Research Methodology* [on line], April 2013, vol. 16, n° 3, p. 231-244. <https://doi.org/10.1080/13645579.2013.774179>

Archives Unleashed Project. *The Archives Unleashed Toolkit* [on line]. Archived on 27 May 2019. <https://web.archive.org/web/20190527054243/https://archivesunleashed.org/aut/>

BANOS, Vangelis; MANOLOPOULOS, Yannis. A quantitative approach to evaluate Website Archivability using the CLEAR+ method. *International Journal on Digital Libraries* [on line], June 2016, vol. 17, n° 2, p. 119-141. <https://doi.org/10.1007/s00799-015-0144-4>

BLANCO, Patricia. *Saving the Belgian Web: Web archiving practices, research opportunities and limitations*. KU Leuven, 2019. Master's thesis. MSc in Digital Humanities.

CHAMBERS, Sally; MECHANT, Peter; GEERAERT, Friedel. Towards a national web archive in a federated country: a Belgian case study. In Brügger, Niels; Laursen, Ditte (eds.) *The Historical Web and Digital Humanities*. New York: Routledge, 2019, p. 29-44.

Internet Archive Blogs. *If you see something, save something* [on line], 1 January 2017. Archived on 16 April 2019. <https://web.archive.org/web/20190416230839/https://blog.archive.org/2017/01/25/see-something-save-something/>

VLASSENROOT, Eveline; CHAMBERS, Sally; DI PRETORO, Emmanuel; GEERAERT, Friedel; HAESENDONCK, Gerald; MICHEL, Alejandra; MECHANT, Peter. Web archives as a data resource for digital scholars. *International Journal of Digital Humanities* [on line], April 2019, vol. 1, n° 1, p. 85-111. <https://doi.org/10.1007/s42803-019-00007-7>

## Notes

1. *PROMISE* stands for PReserving Online Multiple Information: towards a Belgian StratEgy.
2. Knowledge of these languages was also instrumental in filtering the results. The methodology and list of keywords can be found in the student Master's thesis. In References, Blanco, Patricia, *Saving the Belgian Web: web archiving practices, research opportunities and limitations.*
3. French, Dutch, German, English, Italian, Portuguese, Spanish, Galician, Catalan and Basque.
4. Archive Ready <http://archiveready.com/> , tool developed with the *CLEAR+ method*. In References, Banos, Vangelis; Manolopoulos, Yannis. A quantitative approach to evaluate Website Archivability using the CLEAR+ method.
5. Examples of open web archives: Wayback Machine <https://archive.org/web/>, Portuguese Web Archive <https://arquivo.pt/>, UK Web Archive < https://www.webarchive.org.uk/>.
6. For instance, the Right to erasure, or "Right to be forgotten", included in the European General Data Protection Regulation (GDPR) policy.
7. Gephi <https://gephi.org/>
8. Ackland, Robert; Gibson. Hyperlinks and networked communication: a comparative study of political parties online.
9. The text extracted from *La Maison de l'Amérique Latine* (archived on 16 April 2019) <https://web.archive.org/web/20190416014159/https://www.america-latina.be/> was extracted with the Archives Unleashed Toolkit (AUT) and visualized with Voyant Tools <https://voyant-tools.org/>.
10. Internet Archive Blogs. *If you see something, save something.*