A ROADMAP FOR ESTABLISHING A BELGIAN WEB ARCHIVE AT THE FEDERAL LEVEL

Rolande Depoortere, Friedel Geeraert,
Sébastien Soyez, Sophie Vandepontseele

Abstract: The web has become a central part of our daily lives. However, the short lifespan of online data poses serious challenges for preserving and safeguarding this digital heritage. Not preserving web content would leave a considerable gap in the Belgian historical record. To tackle this challenge, web preservation needs to be addressed on the national level as a cultural and historical necessity, which is why the PROMISE research project (2017–2019) was initiated. As Federal Scientific Institutions (FSI) specialised in heritage preservation, the Royal Library of Belgium and the National Archives partnered with the Universities of Ghent and Namur and the university college Bruxelles-Brabant to bring together their expertise in IT, law, information science and digital humanities to study web archiving. The web archiving strategy resulting from this research project is presented in this article comprising the different phases of the web archiving process, the legal framework, a stakeholder analysis and cost calculation for the different web archiving scenarios.

INTRODUCTION

Given the central role that the web plays in society, it contains many traces of our contemporary history, making it an indispensable source for future research. Web content however is notoriously ephemeral, which implies that specific efforts are necessary to preserve this content for future generations. Without such efforts, there is a significant risk of being confronted with a digital black hole[1] and losing a large part of the digital heritage available online.

Many initiatives already exist throughout the world. National libraries have generally taken the initiative to experiment with web archiving. This can be explained by the fact that national libraries aim to collect their country's publishing heritage through legal deposit. In this way, the web is perceived as a large-scale content publisher. Awareness of the need to safeguard this heritage emerged as early as 1996 in Australia, when the National Library

[1] This phrase has been attributed to Vince Cerf, Vice-President and Chief Internet Evangelist for Google, see for example: Dartnell, Lewis. 2015. The digital black hole: will it delete your memories? *The Guardian*, February 16. https://www.theguardian.com/technology/2015/feb/16/digital-black-hole-delete-memories-information-lost-google-vint-cerf (accessed July 8, 2020).

10.1484/J.IMA.5.122153

launched the PANDORA project.[2] The National Library of France also launched its first experimental collections in 1996.[3] Other initiatives were quickly launched in Sweden, New Zealand and the United States with the Library of Congress.[4] The need to bring these different initiatives together, to share knowledge and find joint solutions to the many technological challenges related to the collection of web archives quickly became obvious. It was for this reason that the International Internet Preservation Consortium (IIPC) was created in 2003 to support such web archiving initiatives.[5]

The State Archives and the Royal Library of Belgium (KBR) have taken the initiative to study the possibility of setting up a web archive at the federal level in Belgium. The major challenges in initiating a structural policy for web archiving are the technological constraints and the legal aspects related to copyright and privacy.

The purpose of this article is to present the shared web archiving strategy that was devised by KBR and the Belgian State Archives in the context of the PROMISE research project. The PROMISE project was initiated in 2017 and ran until December 2019, funded by the Belgian Science Policy Office as part of its BRAIN.be programme.[6] Its aim was to develop a federal strategy for the preservation of the Belgian web. KBR and the State Archives partnered with Ghent University (Research Group for Media, Innovation and Communication Technologies and Ghent Centre for Digital Humanities), University of Namur (Research Centre in Information, Law and Society) and Haute Ecole Bruxelles-Brabant (College of Higher Education). The interdisciplinary nature of the project team ensured that the legal, technical and operational aspects, as well as the user needs related to web archiving were investigated.

The concrete objectives of the project were to: 1) identify best practices in the field of web archiving, 2) devise a strategy for archiving the Belgian web, 3) pilot access and use of the Belgian web archive, and 4) make recommendations for the implementation of a sustainable web archiving service. This article focuses on the second objective, namely devising a strategy for archiving the Belgian web. The study of best practices in Belgium and abroad has made

[2] National Library of Australia. 2019. PANDORA Australia's web archive. https://pandora.nla.gov.au/ (accessed July 8, 2020).

[3] Bibliothèque nationale de France. 2020. Archives de l'internet. https://www.bnf.fr/fr/archives-de-lint-ernet (accessed July 8, 2020).

[4] The membership list of the International Internet Preservation Consortium (IIPC) provides a good overview of existing web archiving initiatives worldwide (IIPC 2020). Furthermore, the PROMISE project team identified (interest in) web archiving initiatives at the following institutions: Felixarchief Antwerpen, Universiteitsbibliotheek Gent, Liberaal Archief, AMSAB - Instituut voor Sociale Geschiedenis, ADVN (Archief voor Nationale Bewegingen), KADOC - Documentation and research center on religion, culture and society, Letterenhuis, Archief Gent and Université Catholique de Louvain.

[5] IIPC. 2020. Who is the IIPC? http://netpreserve.org/about-us/ (accessed July 8, 2020).

[6] The acronym PROMISE stands for Preserving Online Multiple Information: towards a Belgian strategy. KBR. 2019. PROMISE project. https://www.kbr.be/en/projects/promise-project/ (accessed July 8, 2020).

it possible to draw lessons from the experiences of institutions that have been dealing with web archiving for some time.

The article starts by providing a brief overview of the relevant legal framework for web archiving that is applicable for the State Archives and KBR. Then, the different steps in the web archiving process will be discussed. In the third section an analysis of the stakeholders who (could) play a role in the Belgian web archive is outlined. The paper concludes with a discussion on the calculations of costs for the different web archiving scenarios that KBR and the State Archives have proposed, followed by the future prospects and a conclusion.

LEGAL FRAMEWORK

When devising a web archiving strategy, it is necessary to consider the legal and organisational framework. Legal mandates provide the boundaries or extent of what can legally be archived, and how, as well as its reuse. The organisational capacity also needs to be considered to assess what is feasible given the expertise and mandates of the organisation, and especially for federal institutions.

In Belgium, and specifically for the two federal institutions involved in the web archiving project we discuss in this article, two Royal decrees are applicable. The activities of KBR are framed by the legal deposit legislation and the royal decree on the establishment of the Royal Library.[7] For the State Archives the most important legislation is the Law on Archives, the royal decree on the establishment of the State Archives of Belgium and the royal decree on executing the Law on Archives.[8] These different legal texts define the scope of the missions of the two institutions. KBR is the national scientific library in Belgium. Its main mission is to collect and to provide access to all Belgian publications and to preserve, manage and study cultural heritage. Web archiving as such is not specifically mentioned in the legal deposit that the Library is mandated to undertake. However, there was a revision to the law in 2018 which stipulates that all publications on digital carriers of all kinds are covered by legal deposit.[9] The royal decree on the establishment of the Royal Library also stipulates that KBR may compile, by means of inventories, lists of websites that are linked

[7] Royal Decree of 19 June 1837. 1837. Arrêté royal du 19 juin 1837 portant constitution en établissement scientifique de la Bibliothèque royale de Belgique. *Moniteur belge*. 8 July 1837.

Law of 8 April 1965. 2018. Loi du 8 avril 1965 instituant le dépôt légal à la Bibliothèque royale de Belgique, telle que modifiée par la loi du 8 juillet 2018. *Moniteur belge.* 20 July 2018.

[8] Law of 24 June 1955. 2009. Loi du 24 juin 1955 relative aux archives, telle que modifiée par la loi du 6 mai 2009. *Moniteur belge*. 19 May 2009; Royal Decree of 18 August 2010. 2010. Arrêté royal du 18 août 2010 portant exécution des articles 1er, 5 et 6bis de la loi du 24 juin 1955 relative aux archives. *Moniteur belge.* 23 September 2010.

[9] Law of 8 April 1965. 2018. Loi du 8 avril 1965 instituant le dépôt légal à la Bibliothèque royale de Belgique, telle que modifiée par la loi du 8 juillet 2018. *Moniteur belge.* 20 July 2018.

to its missions.[10] Thus, with the combination of the obligation of the legal deposit and the revisions to this decree, it can be concluded that websites as digital carriers of information, are included in the deposit.

The State Archives acquire and preserve archives and ensure that they are transferred according to archival standards. The Law on Archives stipulates that archival documents stemming from the Courts of Justice, the Council of State, the Provinces and the public institutions that are subject to their control, or their administrative supervision need to be transferred to the State Archives after a period of 30 years. Archival documents of municipalities and public institutions under their control or administrative supervision on the other hand can also be transferred to the State Archives. Such institutions may also acquire archives of private entities and persons. The royal decree of 18 August 2010 defines archives as all documents that are intended to be preserved by a public authority, private person, company or association governed by private law, regardless of their date, material form, state of preparation or medium, in so far as such documents have been received or produced in the course of their activities or functions. The way in which archives are defined is therefore very important in a web archiving context, as the websites of federal institutions clearly fall under this definition. The legal framework within which the State Archives work was translated into operational selection guidelines (see point 3.1 for detail).

Therefore, according to the archival legal framework, the websites of federal public services must be preserved by the State Archives, with the exception of public services under the legislative bodies. In this way, the boundaries of the Belgian web, that are subject to legal archiving and collection, are defined.

THE WEB ARCHIVING PROCESS

Taking into account the legal boundaries of what can be archived, we developed a workflow to archive the Belgian Web. This workflow is based on the OAIS-model[11] and its six functional entities: ingest, data management, archival storage, access, preservation planning and administration (ISO 14721:2012). When archives are ready to be transferred to an archiving system, they are first ingested into the system. Secondly, the descriptive and technical metadata need to be properly managed by a specific step; the data management function. In parallel to this, the physical preservation of the archives is made possible through storage facilities. Furthermore, the preservation sustainability must be organised within a dedicated preservation planning process, so the formats and storage medium's obsolescence are properly managed. In order to design and enable consultation of the archives, the access

[10] Royal Decree of 25 December 2016. 2017. Arrêté royal modifiant l'arrêté royal du 19 juin 1837 portant constitution en établissement scientifique de la Bibliothèque royale de . *Moniteur belge.* 16 January 2017.

[11] The acronym OAIS stands for Open Archival Information System. OAIS. 2020. OAIS Reference Model (ISO 14721). The fundamental standard for digital preservation. http://www.oais.info (accessed July 8, 2020).

function needs also to be set-up, so the user can search and retrieve the information he/she needs, in an appropriate way (e.g. technically, legally). Finally, a general administration function is also foreseen in the OAIS model. The purpose of this function is to ensure an operational responsibility around the above-mentioned functions, like a conductor should have with his or her orchestra. Based on these 6 functions, and due to the specificity of the web archives, we decided to add three additional phases to the web archiving workflow namely selection, capture and quality control. The objective was to describe in detail the first steps of the archival process, namely the selection criteria and the technical steps of the capture. The quality control function was added just before ingest, in order to identify any technical issues. Thus, a double selection strategy for the Belgian web archive was devised by the State Archives and KBR to consider these two different, but overlapping, mandates.

Selection

Within the PROMISE project an operational definition of the Belgian web was devised based on the definitions of a national web in the legal deposit legislation in France[12] and Denmark,[13] amongst others. The Belgian web comprises three different categories of web content. Firstly, content on domains that have a link with the Belgian territory, such as the national domain, and domains for cities, regions, e.g. .be, .brussels, .vlaanderen and .gent. Secondly, web content of other country code Top Level Domains (ccTLDs), e.g. .fr, .de, .uk, or general Top Level Domains (gTLDs) which include .org, .com or .net domains. For this second category, content can be considered as Belgian as long as these domain names have been registered by Belgians or concern Belgium from a historical, political, or cultural perspective, as well as web content that is of interest or relevant to the country or to Belgian society at large which includes, but is not limited, to national cultural heritage. Thirdly, websites of which part of the activities linked to the creation, production or publication of web content took place on the national territory, are also eligible. On an operational level, the last criterion can be translated into identifying which websites are hosted in Belgium, using Geo-IP localisation tools. As each country is assigned a range of IP-addresses, these tools can detect whether a certain website is hosted on a server located in Belgium. Geo-IP localisation is not an exact science however, and estimating the number of domain names that fall within the latter category is, as a consequence, not as precise as we expect.

In September 2019, the Belgian web included 1,6 million .be domain names, 3,400 .gent domain names, 7,800 .brussels domain names and 6,400 .vlaanderen domain names. It also needs to be noted that these figures include domain names that are linked to non-active domains or to mail boxes. In this way, the number of active domain names linked to a

[12] Bibliothèque nationale de France. 2020. Archives de l'internet. https://www.bnf.fr/fr/archives-de-linternet (accessed July 8, 2020).

[13] Danish Royal Library. 2020. Netarchive.dk collects and preserves the Danish part of the internet. http://netarkivet.dk/in-english/ (accessed July 8, 2020).

website is likely to be lower than these figures. Accordingly, based on this volume, we can estimate that approximately 1 million websites are potentially active. Still, it is clear that there is a lot of material to be archived.

Given the double selection strategy for the Belgian web archive, we developed two scenarios for crawling the web. On the one hand a broad crawl consisting of the top level domain names pertaining to the Belgian web as defined above, and on the other hand selective crawls focusing on specific themes or events. The broad crawl would be limited to the top layers of the websites and would ideally be crawled once a year, whereas the websites in the selective collections would be crawled completely and more frequently than once a year. The issue with the selection of content in the context of web archiving is that it is very difficult to anticipate what will be of interest to future users. A combination of both strategies would therefore allow a broad sample of the Belgian web to be preserved as well as more in-depth collections about specific themes and events. The selective collections of the State Archives and KBR differ given the specific missions entrusted to both institutions, as detailed in the previous section. This resulted in two different test crawls. A test crawl of the following websites was undertaken within the PROMISE project following the operational selection guidelines of the State Archives:

1. websites of federal institutions: executive, judicial and legislative powers;
2. websites of ministerial cabinets, ministers/secretaries of state;
3. websites of other public organisations that have a link with the federal level: trade associations, trade unions, federations, public health insurance organisations, political parties, public interest organisations, the monarchy, etc.; and
4. websites of the provinces, the regions and the communities.

This resulted in a complete seed list, or list of URLs to archive, comprising of around 650 websites. In the future, the list could be expanded by adding websites of cities, municipalities and other public institutions (federal, regional or local) and websites linked to private archives acquired by the State Archives.

The themes that have been identified for selection at KBR represent the various departments and thus collections of KBR. For example, this could include websites and thus topics related to specific collections of music, prints, manuscripts, coins and medals, restoration and conservation in Belgium, etc. For the Department of contemporary collections, themes such as Belgian literary blogs, publishers, e-magazines, comic books or (youth) literature can be included. Other specific collections could also be curated around specific events or themes that play an important role in society, such as the representation of minorities on the web. Within the PROMISE project a number of these collections were created.[14] The full seed list comprised 920 websites and 1,400 web pages.

[14] KBR. 2019. PROMISE project. https://www.kbr.be/en/projects/promise-project/ (accessed July 8, 2020).

Capture and Quality Control

Capturing web content is done by means of a web crawler. Web crawlers identify and follow the hyperlinks on web pages, capturing and saving the information they encounter along the way.

Several tools for capturing the web exist, but within the framework of the project, it was chosen to use the one that is the most widely known[15] to test the collecting process. It starts with one of the domain names included on the seed list and then follows the internal links between all the pages of a website. Heritrix makes a copy of every single page of a website and stores it into a file called a WARC (Web ARChive) file. The WARC file format is an ISO standard (ISO 28500:2017) and the most used file format within web archiving.

Several crawler parameters can be set, for example whether or not the crawler should stay within the website or may also collect websites to which the initially selected website points. The depth and frequency of a crawl can also be set. The depth of a crawl is mostly expressed in the number of clicks or hops the crawler can take from the homepage. The crawl frequency highly depends on the collection. The websites of Belgian e-magazines or e-newspapers would logically be captured more frequently than content that does not change as frequently.

It is a misconception that the crawling process results in a perfect capture of the content that was online at that moment. Dynamic content for example is notoriously difficult to capture and elements such as stylesheets are frequently not captured by the crawler, resulting in a capture that looks very different from the original. Incompleteness therefore needs to be considered as an inherent trait of web archiving. Since no automated tools exist yet for quality control, manual quality control was tested within the PROMISE project. However, it proved to be very labour intensive and was thus not undertaken beyond this initial test. The answer to the question of how quality control will be organised in the future Belgian web archive will therefore largely depend on the available resources that can be allocated to this task. In general, the selective collections would merit a higher level of quality control than the broad crawl given the sheer size of the latter. Performing quality control on a limited sample of websites is also a potential approach.

Ingest

After the websites have been crawled and have undergone quality control, the resulting WARC files need to be ingested into the systems of KBR and the State Archives. A number of necessary checks need to be done before the importing these files, which include anti-virus

[15] Internet Archive. 2018. Heritrix 3 documentation. https://heritrix.readthedocs.io/en/latest/ (accessed July 8, 2020).

checks, format validation and integrity checks, by means of checksums. The State Archives and KBR both work with the METS container format,[16] thus the use of this format is also a requirement for the future Belgian web archive.

The frequency with which the WARC files are ingested depends on the frequency of crawling. Collections that are captured frequently can be ingested on a regular basis. Thus, this step is dependent on the websites crawled and how often they may change, as well as the mandates of the institutions to document these changes. In the PROMISE project, one of the strategic goals of KBR was to make information available to the public as quickly as possible, thus regular ingests are therefore a requirement for KBR. For the State Archives, one ingest per year would in theory suffice as the institution would crawl their seed list at minimum once a year.

Storage

Maintaining and storing these files is an important element for the institutions. In the PROMISE project, we recommended that the WARC files will be stored as determined in the disaster recovery plan of both institutions for access and preservation. Hashing algorithms will be used to perform the necessary integrity checks. This method allows to "calculate" by a hash-function the content of a file, and later on, to compare it with a new calculation of the same file. If the two calculations match, then the 2 files are identical. If not, it means that there was a modification (even if slight) in the file and, consequently, its integrity cannot be guaranteed. For long-term preservation both institutions can use the LTP platform managed by BELSPO to safely store their web archive collections. Operational procedures for this data transfer to the LTP platform still need to be drafted and tested.

It is important to mention that the cost for storage can be reduced by making use of deduplication and compression. Given that this approach was not tested within the PROMISE project, operational procedures for this point still need to be drafted.

Data Management

Different kinds of metadata need to be preserved in the web archiving process so that the files can be catalogued and categorised within the institutions. Technical metadata are automatically stored in the WARC file. A WARC file consists of multiple WARC records, one of which is dedicated to metadata. For descriptive metadata, the State Archives and KBR plan to make use of the OCLC model for descriptive metadata for web archiving,[17]

[16] Library of Congress. 2019. METS Metadata Encoding and Transmission Standard. http://www.loc.gov/standards/mets/ (accessed July 8, 2020).

[17] Dooley, J. & Bowers, K. 2018. Descriptive metadata for web archiving. Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group. Technical report. https://www.oclc.org/research/publications/2018/oclcresearch-descriptive-metadata.html (accessed July 8, 2020).

because it provides a standard set of elements that are useful for tracing and categorising the websites. The choice of OCLC was mainly made because it is a standard internationally supported and promoted by the library community. Moreover, the OCLC Web Archiving Metadata offers an easy link to both MARC21 and EAD, the metadata standards used by KBR and State Archives respectively. Concretely, the dataset comprises of 14 metadata elements: URL, title, creator, contributor, language, collector, date, subject, genre/form, relation, description, extent, rights and source of description. Using the same metadata set would allow KBR and the State Archives to make use of a shared access platform. It is important to note that the creation of elaborate descriptive metadata is not possible for the broad crawl given the fact that manual metadata creation is very time-consuming. For the broad crawl, the descriptive metadata would be limited to those elements that can be automatically created based on the information provided by the website itself such as URL, language and title.

Preservation Planning[18]

WARC files are complex digital objects as WARC is a container file format that encapsulates all the content contained within a website, with all the variety in data formats this implies. Currently there are very few recommendations regarding the conservation of WARC files. Since the preservation processes were not tested within the PROMISE project, the operational procedures still need to be drafted. Tools such as JHOVE[19] or DROID[20] can be used to ensure that the WARC file complies with the WARC ISO standard (ISO 28500:2017).

It is essential to be able to identify the files contained in a WARC file in order to detect whether any risks are associated with any of the file formats. If any risks are detected, migrating the content to another format is necessary. In this case, the WARC file needs to be decompressed, the problematic file needs to be migrated and the WARC file needs to be recompressed. The integrity of the recompressed WARC file needs to be maintained by updating all the links that linked the migrated files to other elements of the web page. The conversions also need to be noted in a conversion register to ensure that the changes can be traced.

[18] KBR is currently working on a preservation needs analysis for all its digital content, including web archiving. Virginie Rodriguez is also preparing an article for the next issue of the journal "Les Cahiers de la Documentation."

[19] JHOVE is Open source file format identification, validation & characterisation a is a format-specific digital object validation. Open Preservation Foundation. 2015. JHOVE Open source file format identification, validation & characterisation. http://jhove.openpreservation.org/ (accessed July 8, 2020).

[20] DROID is a software tool developed by The National Archives to perform automated batch identification of file formats. The National Archives. 2020. Download DROID: file format identification tool. https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/ (accessed July 8, 2020).

Access

KBR and the State Archives aim to make the web archive collections available in multiple ways. Access to web archives is considered to be one of the biggest challenges. This is largely due to copyright legislation, which limits access to archived web materials to reading rooms in many countries. Unfortunately, this is also the case in Belgium.

One of the ways in which KBR and the State Archives wish to provide access to the collections is via their respective catalogues. As indicated above, descriptions of the selective collections in EAD and MARC21 can be created based on the descriptive metadata proposed in the OCLC study. A second way to access the collections is a shared access platform on which collections can be viewed via replay software. Replay software allows the user to interact with an archived website much like he or she would with a website on the live web. Several kinds of replay software exist but within the PROMISE project PyWB was used to pilot access to the collections (Python Software Foundation 2020.)[21] Ideally the access platform should offer elaborate search methods: browsing (alphabetically, per collection, per subject, etc.), advanced search options (word in title, full-text search, URL-search, image search, etc.) and filters for file type, collection, language, subject, data, etc. Given that the access platform currently exists as a basic prototype, the available functionalities on the future access platform will largely depend on the resources that are available at the time of development.

It is also important to stay informed about the user needs within the context of web archives. More and more researchers approach web archives as corpora for big data analysis. This implies that offering simple replay software that allows interacting with a single website will not be sufficient for this user profile. Offering direct access to the WARC files would be a solution to the problem, however, certain conditions apply from a legal perspective. Access to WARC files should be restricted to academic researchers and the legitimate interest in the requested WARC files should be evaluated based on a motivated request by the researchers. In addition, a user agreement detailing what kinds of use are permitted also needs to be drawn-up. Furthermore, researchers may only make use of the WARC files in the reading rooms of the State Archives and KBR.

Researchers are also interested in documentation about curatorial decisions and the crawler settings that were used. Providing access to documents such as crawl logs or access logs is essential for researchers wishing to gain more insight into the crawling process since the crawler settings have a direct influence on the web archive data that are made available to the public. Making available additional metadata related to the crawls i.e. the depth, frequency, duration, quality assurance criteria, IP address of the server, comments, and so forth, on the access platform would further improve this access.

[21] Python Software Foundation. 2020. pywb 2.3.5. https://pypi.org/project/pywb/ (accessed July 8, 2020).

Another way to encourage research use of web archive data is by making available derivative datasets. Derivative files can, for example, contain text extracted from the HTML pages of a number of web pages, an export of technical metadata such as the seed list and the corresponding timestamps of crawling, or an export of all the hyperlinks included in (a specific section of) a web archive. The legal framework comes into play again as copyright legislation applies. In case of the examples mentioned, the text extracted from HTML pages is usually protected by copyright and these derivative files should therefore only be made available in the reading room. In the other two cases, it concerns either metadata that was created by KBR and the State Archives, or elements of websites such as hyperlinks which are not protected under copyright law. These derivative files could in theory be made freely accessible online.

Administration

Administration should be interpreted as the operational management of the web archive. It is clear that a lot of stakeholders are involved in web archiving activities and that coordination of these different activities is required. In the following section we detail who these stakeholders are in the stakeholder analysis for a functional Belgian web archive.

On the human resources level, web archiving teams usually comprise of two different profiles: technical (crawl engineer, (software) developer, IT specialist, programmer, etc.) and curatorial / archival / librarian (web archivist, digital curator, etc.). The tasks related to administration are very diverse which include: providing training in selection of content and creation of descriptive metadata, installing and maintaining the necessary tools and servers, managing relations with external parties and stakeholders, communicating about the web archive and organising outreach, following-up of complaints, drafting and updating technical procedures, documenting problems that have been identified, monitoring and reporting, and defining technical and functional requirements of the web archiving infrastructure. Thus, institutions have to consider and prioritise different kinds of expertise in maintaining a web archive.

Strategic Management

The strategic management teams within both institutions are responsible for evaluating the web archiving strategy and the collection development plans. Other strategic decisions are linked to, for example, the management of changes in the applicable legislation, risk analysis and management or making decisions about memberships of (international) organisations related to web archiving. It is up to the Board of Directors of each institution to take full responsibility for these strategic issues. Of course, the political choices made on an higher level can influence this strategy positively or negatively.

STAKEHOLDERS

Setting up a Belgian web archive involves many actors who play either a direct or indirect role. In this section, the main stakeholders of the Belgian web archive are presented.

KBR and the Belgian State Archives can be considered as the focal organisations since the initiative for setting up a Belgian web archive stems from these two institutions. Belgian heritage institutions can play a significant role as well. Knowledge and expertise can, for example, be exchanged with other Belgian organisations who are active in the field of web archiving.[22] Another option is for heritage institutions to submit suggestions for content to be harvested by KBR and the State Archives in order to allow for an inclusive selection policy or, taking it one step further, curate their own collections of websites and contract KBR and the State Archives as service providers for web archiving. Registries, the bodies that manage certain domains, are also stakeholders in the project because they can deliver exhaustive lists of domain names pertaining to Belgian domains such as .be, .vlaanderen, .gent and .brussels. ICANN[23] can provide lists for .gent, .vlaanderen and .brussels and DNS Belgium[24] can, in theory, provide the list for the .be domain.

Depending on the chosen strategy another important partner for the Belgian web archive could be an external service provider. A number of organisations offer web archiving services for a fee, such as Archive-It[25] and the Internet Archive.[26] More concretely, this would mean that KBR and the State Archives select the content and send the seed lists to the service provider. The service provider would crawl, make the content available and send a copy of the content to KBR and the State Archives.

A further important group of stakeholders are the users of the web archive. Web archives generally contain a wealth of information that can interest a large variety of people including the general public or researchers such as political and social scientists, digital humanists, historians, media and communication scientists and linguists, etc. It is very important to take into account the user needs when devising a web archiving strategy. Belgian society at large is also key for the web archive as it produces the content that can be found in the web archive. At the same time, Belgian society constitutes a large number of potential users of the web archive.

On the federal level BELSPO is an important stakeholder in the project because they offer a long-term preservation platform on which preservation copies of the Belgian web archive

[22] A number of Belgian heritage institutions are already involved in web archiving such as Felixarchief Antwerpen, University Library Ghent, University of Antwerp, Liberas, AMSAB – Institute for Social history, Archive for National Movements, KADOC (Documentation and Research Center on Religion, Culture and Society), Letterenhuis, Ghent City Archive.

[23] ICANN. 2020. Internet Corporation for Assigned Names and Numbers. https://www.icann.org/ (accessed July 8, 2020).

[24] DNS Belgium. 2020. Registry for .be, .brussels and .vlaanderen. https://www.dnsbelgium.be/en (accessed July 8, 2020).

[25] Archive-It. 2020. Web archiving services for libraries and archives. https://archive-it.org/ (retrieved July 8, 2020).

[26] Internet Archive. 2020. Internet Archive. https://archive.org/ (accessed July 8, 2020).

can be stored. Moreover, government bodies who are subject to the Law on Archives, are important stakeholders since their content can be included in the part of the web archive managed by the State Archives. Especially the bodies hosting the websites of the Belgian government (Federal Public Service (*FPS*) Policy and Support (BOSA), Smals and FPS Chancellery of the Prime Minister) can be key partners since they have a good view on which websites are currently in use. The Minister in charge of the Digital Agenda and the Minister or State Secretary in charge of Science Policy are other stakeholders because they regulate the legal framework within which KBR and the State Archives work.

On an international level a number of organisations exist that are interesting partners. The International Internet Preservation Consortium (IIPC) is a partner in the web archive because it offers an international platform for knowledge exchange and standard development within the field of web archiving (IIPC 2019.)[27] KBR is an IIPC member since 2018. RESAW, the Research Infrastructure for the Study of Archived Web Materials, is also a stakeholder because it offers an international platform for the valorisation of research results based on the content included in the Belgian web archive (RESAW 2019.)[28] Additionally, the WARCNET network[29] is a new international initiative that promotes national and transnational research to study the history of the transnational web and therefore also constitutes an interesting stakeholder.

COST CALCULATION FOR DIFFERENT WEB ARCHIVING SCENARIOS

This section provides more information about the detailed cost calculations that were undertaken within the PROMISE project, based on three particular scenarios for collaboration between the State Archives and KBR.

Scenarios

KBR and the State Archives worked on different web archiving scenarios to present to their respective Management Boards. This allowed to offer a variety of possible institutional approaches to web archiving depending on the resources that can be made available. These scenarios were based on differences in selection and span three different levels: full, medium and basic.

The full scenario covers selective collections and a broad crawl comprising of 100% of the Belgian web. The selective collections for the State Archives would comprise of (1) the

[27] IIPC. 2020. International Internet Preservation Consortium. http://netpreserve.org/ (accessed July 8, 2020).

[28] RESAW. 2019. A Research Infrastructure for the Study of Archived Web Materials. https://resaw.eu/ (accessed July 8, 2020).

[29] WARCNET. 2020. Web ARChive studies network researching web domains and events. https://cc.au.dk/en/warcnet/ (accessed July 8, 2020).

websites of the federal institutions the State Archives are legally obliged to preserve (option A.1), (2) the websites of cities, municipalities and other (federal, regional and local) public institutions (option A.2), and (3) websites from private archives (option A.3). For KBR the selective collections would comprise of websites that are closely linked to the existing collections and fit within the collection development plan (option A.1). The websites that are part of the selective collections would be completely captured. For the broad crawl, the crawler would be limited to only crawling the top layers of all websites pertaining to the Belgian web, thereby constituting a sample (option B.3).

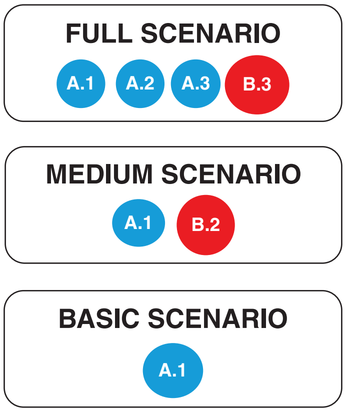
In the medium scenario, the broad crawl would be limited to a randomly chosen sample of 10% of the Belgian web (option B.2). The selective crawls for the State Archives would be limited to the websites of the federal institutions of which the State Archives are legally obliged to preserve the archives (option A.1). For KBR the selective collections would cover the same content as in the full scenario (option A.1). The basic scenario includes the same selective collections as in the medium scenario, but the broad crawl is excluded.

Methodology and Calculation

A second step in the process was to calculate the cost for each of these scenarios assuming that KBR and the State Archives would collaborate on developing and maintaining the necessary infrastructure. To this end, a comprehensive and in-depth analysis was done in which all tasks within the different steps of the web archiving process were listed. Estimates were made of the number of hours required per task. A distinction was made between recurrent and one-off tasks. Each task was also assigned to the relevant job profiles, which allowed the approximate labour cost to be calculated for each task. In addition to the cost for human resources, the cost of the web archiving infrastructure was also calculated. Estimates of the size of the various collections were made based on input from the ICT services of the State Archives and KBR and the researchers specialised in ICT within the PROMISE project. The annual cost of human resources and ICT infrastructure was calculated over a period of five years, which was then divided by five in order to calculate the average annual cost for the first five years of a functional web archive on the federal level. Table 1 below, outlines the annual cost per scenario based on a jointly managed web archive for the State Archives and KBR.

Analysis

In the full scenario, the selective crawl covers the entirety of 2,350 websites for the State Archives and of 920 sites for KBR, plus 1,400 pages from other sites for KBR. The broad crawl includes a partial capture of each of the million active Belgian websites. This scenario requires a staff of 5.76 FTE (about 275,000 euros). The update of the seed list, the selective crawl (including quality control) and the operational management are the most demand-



FULL SCENARIO							
	AGR			KBR			AGR + KBR
TOTAL	175.884 €			303.222 €			479.106 €
STAFF	2,5 FTE	125.272 €	71%	3,3 FTE	148.568 €	49%	273.840 €
INFRASTRUCTURE	50.612 €		29%	154.654 €		51%	205.266 €
MEDIUM SCENARIO							
MEDIUM	AGR			KBR			AGR + KBR
TOTAL	85.599 €			273.016 €			358.615 €
STAFF	1,5 FTE	73.583 €	86%	3,3 FTE	148.568 €	54%	222.151 €
INFRASTRUCTURE	12.016 €		14%	124.448 €		46%	136.464 €
BASIC SCENARIO							
BASIC	AGR			KBR			AGR + KBR
TOTAL	82.243 €			269.660 €			351.903 €
STAFF	1,5 FTE	73.583 €	89%	3,3 FTE	148.568 €	55%	222.151 €
INFRASTRUCTURE	8.660 €		11%	121.092 €		45%	129.752 €

Table 1. Average annual cost of the different web archiving scenarios

ing tasks where staff is concerned, as the broad crawl is a fully-automated process without significant human input. The volume of the capture is estimated at 114 TB for the first year (40 TB resulting from the broad crawl, 70 TB from the selective crawl for KBR and 4 TB from the selective crawl for the State Archives). The high proportion of the selective crawl for KBR can be explained by the nature and the frequency of the crawl. Some press websites are composed of numerous large files (videos, sound, images) and will be crawled several times a day because of their highly dynamic content. In comparison, the selective crawl of 2,350 sites for the State Archives weighs only 4 TB because these sites are less frequently modified and a crawl once a year is considered to be sufficient. This is similar to the collections of other digital archives from the public services concerned. An annual growth of 10% is integrated in the calculation. The implementation of the infrastructure would cost about 205,000 euros. Overall, the full scenario comes to about 480,000 euros annually.

In the medium scenario, the number of websites to be crawled in their entirety is considerably reduced for the State Archives (650 domain names) and stays identical for KBR, while the broad crawl is reduced to 10%. The total annual weight to be stored decreases to 75 TB and the infrastructure cost is reduced to 136,500 euros. This last reduction is only due to the impact on the storage capacity because the rest of the hard and software stay identical to the full scenario. The cost of staff decreases by 1 FTE due to the fact that the selective crawl for the State Archives is severely reduced and that tasks such as selection and description of websites require less time. The total cost of this scenario is estimated at about 360,000 euros.

In the third basic scenario, the selective crawl is the same as in the medium scenario but there is no broad crawl. This does not significantly lower the costs. The staffing requirements stay identical since the broad crawl does not imply a lot of human intervention and its absence has no influence on the human resources. The storage is reduced by only 4 TB. The basic scenario saves scarcely 6,700 euros compared with the medium one.

The difference in costs between the two institutions is mainly due to the difference between KBR's and the State Archive's selection policies for selective collections. At KBR, the selective collections (option A.1) were estimated to comprise 920 websites and 1,400 web pages, some of which will be captured more frequently than once a year (e.g. newspaper homepages). For option A.1 for the State Archives, there are 650 websites to be archived annually. The more sites selected for selective collections, the higher the cost for the selection and creation of metadata, but also for the updating of the seed list, the quality control and the administrative follow-up of the authorisations of the rights holders of a site for collection and access.

When Option B.3 is considered, we note that the costs of the State Archives increase much more than at KBR. Infrastructure costs increase by about 33,000 euros for each institution because of the addition of option B.3 (50% - 50% split). Furthermore, the addition of options A.2 and A.3 for the State of Archive incurs additional staff costs because of the selection and creation of metadata, the annual update of the seed list, quality control and the follow-up of the authorisations of the right holders of websites from private archives. At KBR, staff costs remain stable between scenarios 2 (medium) and 1 (full) because there is no additional manual seed selection, metadata creation or quality control since the selective collections are the same in both scenarios.

FUTURE PROSPECTS AND CONCLUSION

The PROMISE research project is a fundamental step towards the implementation of a structural web archiving policy at the federal level for Belgium. Currently, the Belgian Web Archive only exists as a prototype, however, the aim is to develop a functional Belgian Web Archive in the coming years. The concrete scenarios and the associated costs can serve as a basis for KBR and the State Archives to make further informed strategic decisions about web archiving. Obviously, the available means, both financial and in term of human resources, will be taken into account in these decisions. The conclusions of this project also provide a solid evidence-base to advocate for the necessity of a structural budget dedicated to the management and preservation of and access to digital archives for citizens and researchers, whether they were digitized or were born digitally.

Furthermore, in order to guarantee a coherent and efficient selection policy for the entire Belgian web, it will be necessary to build a model of participative collaboration with the various Belgian heritage institutions in order to allow them to exchange information about existing web archiving practices, gather suggestions for collections or to help them as a service provider with their own web archiving policy.

Today, Belgium has already lost a large part of its historic web due to the lack of it being archived, and thus it is necessary to do everything possible to prevent further content from being irretrievably lost in the future. The web provides records of memories for the public, and as heritage institutions we must ensure that these essential traces of human activity are preserved in order to make them available and accessible to all, both for the citizen who wishes to find a particular website, and for researchers who wish to work on big data analyses based on archived web content.

BIBLIOGRAPHY

Archive-It. 2020. Archive-It – Web archiving services for libraries and archives. https://archive-it.org/.

Bibliothèque nationale de France. 2020. Archives de l'internet. https://www.bnf.fr/fr/archives-de-linternet.

Danish Royal Library. 2020. Netarchive.dk collects and preserves the Danish part of the internet http://netarkivet.dk/in-english/.

Dartnell, Lewis. 2015. The digital black hole: will it delete your memories? *The Guardian*, February 16. https://www.theguardian.com/technology/2015/feb/16/digital-black-hole-delete-memories-information-lost-google-vint-cerf.

DNS Belgium. 2020. Registry for .be, .brussels and .vlaanderen. https://www.dnsbelgium.be/en.

Dooley, J. & Bowers, K. 2018. Descriptive metadata for web archiving. Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group. Technical report. https://www.oclc.org/research/publications/2018/oclcresearch-descriptive-metadata.html.

ICANN. 2020. Internet Corporation for Assigned Names and Numbers. https://www.icann.org/.

IIPC. 2020. Who is the IIPC? http://netpreserve.org/about-us/.

IIPC. 2020. International Internet Preservation Consortium. http://netpreserve.org/.

IIPC. 2020. IIPC Members. http://netpreserve.org/about-us/members/.

International Organisation for Standardisation. 2012. ISO 14721:2012 Space data and information transfer systems – Open archival information system (OAIS) – Reference model.

International Organisation for Standardisation. 2017. ISO 28500:2017 Information and documentation - WARC file format.

Internet Archive. 2018. Heritrix 3 documentation. https://heritrix.readthedocs.io/en/latest/.

Internet Archive. 2020. Internet Archive. https://archive.org/.

KBR. 2019. PROMISE project. https://www.kbr.be/en/projects/promise-project/.

Library of Congress. 2019. METS Metadata Encoding and Transmission Standard. http://www.loc.gov/standards/mets/.

National Library of Australia. 2019. PANDORA Australia's web archive. https://pandora.nla.gov.au/.

OAIS. 2020. OAIS Reference Model (ISO 14721). The fundamental standard for digital preservation. http://www.oais.info/.

Open Preservation Foundation. 2015. JHOVE Open source file format identification, validation & characterisation. http://jhove.openpreservation.org/.

Python Software Foundation. 2020. pywb 2.3.5. https://pypi.org/project/pywb/.

RESAW. 2019. RESAW. A Research Infrastructure for the Study of Archived Web Materials. https://resaw.eu/.

The National Archives. 2020. Download DROID: file format identification tool. https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/.

WARCNET. 2020. Web ARChive studies network researching web domains and events. https://cc.au.dk/en/warcnet/.