Behind the Scenes of Web Archiving: Metadata of Harvested Websites

Emmanuel Di Pretoro, Friedel Geeraert

BEHIND THE SCENES OF WEB ARCHIVING METADATA OF HARVESTED WEBSITES

Emmanuel DI PRETORO, Friedel GEERAERT, Sébastien SOYEZ

Introduction

The web is fraught with contradiction. On the one hand, the web has become a central means of communication in everyday life and therefore holds a lot of sources. Yet, much less importance is attached to its preservation. Online content has a very short lifespan, which means that every day interesting sources for future research are lost. Web archiving initiatives aim to capture this content and preserve it. Web archiving is defined by the International Internet Preservation Consortium (IIPC) as *the process of collecting portions of the World Wide Web, preserving the collections in an archival format, and then serving the archives for access and use*.[1]

The first web archiving initiatives saw the light in the mid-1990s. Library and Archives Canada experimented with archiving web content as part of the Electronic Publications Pilot Project in 1994-1995.[2] The Internet Archive started their web archive in 1996 with a very international collection scope. Several initiatives with a national focus followed and by the year 2000, web archives had been founded in Australia, the UK, Sweden, New Zealand, the USA and the Czech Republic.[3] Nowadays, most countries in Europe have a web archive at national level, the majority of which are managed by national libraries (mostly as part of the legal deposit), but in some countries such as the UK and The Netherlands both the national archives and the national library are involved in web archiving.

A number of web archiving initiatives exist in Belgium, for example at the Felixarchief[4], the University Library in Ghent[5], Liberaal Archief[6], AMSAB[7],

---

[1] IIPC, *Why archive the web?*, 2018. Retrieved from http://netpreserve.org/web-archiving/. Last accessed on 28/03/2019.

[2] Library and Archives Canada, *Electronic Publications Pilot Project (EPPP) Final project*, 1996. Retrieved from http://www.nlc-bnc.ca/obj/p4/f2/e-report.pdf. Last accessed on 09/04/2019.

[3] R. Schroeder and N. Brügger, "Introduction: the web as history", in: N. Brügger and R. Schroeder (Eds.), *The web as history. Using web archives to understand the past and present*, London, 2017, p. 7.

[4] https://felixarchief.antwerpen.be.

[5] https://lib.ugent.be.

[6] https://stad.gent/openingsuren-adressen/liberaal-archiefliberas.

ADVN[8] or Archief Gent[9], but their collection scope is limited. The PROMISE project (PReserving Online Multiple Information: towards a Belgian StratEgy) was therefore initiated by the Royal Library and the State Archives of Belgium in 2017. They partnered with the universities of Ghent (Research Group for Media, Innovation and Communication Technologies; Ghent Centre for Digital Humanities) and Namur (Research Centre in Information, Law and Society) and the university college Bruxelles-Brabant (Unité de Recherche et de Formation en Sciences de l'Information et de la Documentation). The PROMISE project is financed by the Belgian Science Policy Office (BELSPO) as part of the BRAIN.be programme and runs until December 2019.[10]

The goals of the project are 1) to identify best practices in the field of web archiving, 2) to develop a strategy for archiving the Belgian web, 3) to develop a pilot web archive at federal level and 4) to write recommendations for the implementation of a sustainable web archive in Belgium. An essential part of the pilot project is metadata management. The PROMISE project is particularly innovative in the sense that the State Archives and the Royal Library are working together on a national web archive. In most countries where both the national library and national archives are involved in web archiving, the institutions each work on their own web archive.[11] One of the aims of the PROMISE project is to allow as much collaboration and interoperability between both institutions as possible, including metadata management.

This paper first provides more information about web archiving from a technical point of view before focusing on descriptive metadata in the context of web archiving and the WARC file format. Lastly, the experiments done within the PROMISE project with regard to integrating metadata into the WARC file format are discussed.

How is web archiving done?

Before delving further into metadata in a web archiving context, it is important to explain how web archiving is done. There are different ways, but the most common is by means of crawling or harvesting (client-side web archiving). This entails that a crawler robot is fed a list of URLs. The robot sends HTTP requests

---

[7] http://www.amsab.be.

[8] https://advn.be.

[9] https://www.ugent.be/nl/univgent/voorzieningen/collecties/archief.

[10] Belgian Science Policy Office, *Belgian research action through interdisciplinary networks*, 2019. Retrieved from https://www.belspo.be/belspo/brain-be/index_en.stm. Last accessed on 09/04/2019.

[11] One exception is Canada where the national library and the national archives form one single institution (Library and Archives Canada), but as the web is considered to be a published resource, web archiving activities are managed by the National Library side of the organisation. N. Villeneuve and T Smyth. (2018, February 7). Interview with Nathalie Villeneuve & Tom Smyth/ Interviewers: Emmanuel Di Pretoro, Friedel Geeraert, Gerald Haesendonck, Alejandra Michel and Eveline Vlassenroot [flv file].

to a web server and captures the content that is communicated by the server. Simultaneously, the crawler also creates a list of all hyperlinks that are contained in the crawled web pages. Another method is server-side web archiving which requires direct access to the server and therefore also direct participation of partner organisations. A copy of the information on the server is in this case obtained without going through the HTTP protocol. A third possibility is via transaction archiving which is essentially a screen capture that is made of a user accessing and interacting with the content on a particular website.

Within the PROMISE project, client-side web archiving was chosen as this approach allows to capture a large number of websites. While assessing the state of the art, several institutions involved in web archiving were interviewed. One of the questions was about the crawler the institutions in question use.[12] Heritrix is used by more than half of the institutions, or even more than half if one includes the institutions that choose to delegate the capture of online content to service providers (MirrorWeb, Archive-It) as they are also using Heritrix. Different crawlers were tested (wget, Brozzler, WebRecorder, etc.), but Heritrix (version 3.3.0 LBS) was finally chosen for the PROMISE project.

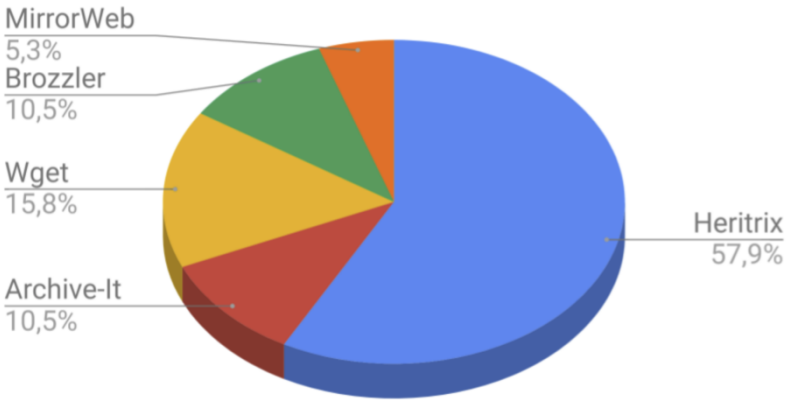

Figure 1: Overview of crawlers used by studied web archiving initiatives

Metadata in web archives

Metadata are strongly linked to the concepts of integrity and authenticity of online information. ISO 15489-1 on "Information and documentation - records

[12] The surveyed institutions are: National Library of Switzerland, Library and Archives Canada, National Library and Archives Québec, National Library of Ireland, Archivo.pt (Portugal), Royal Library of Denmark, British Library, UK National Archives, National Library of Luxembourg, National Library of France and the Institut National de l'Audiovisuel (France), National Library of The Netherlands and National Archive of the Netherlands.

management" defines an authentic record as "one that can be proven to: a) be what it purports to be; b) have been created or sent by the agent purported to have created or sent it; and c) have been created or sent when purported". According to the same standard, "a record that has integrity is one that is complete and unaltered."[13] As is the case for other electronic records, provenance information plays an important role in web archiving. The information about the source of records in a web archiving context includes the URL, content producers, date(s) of creation/transmission, etc.[14]

As web archiving via client-side web archiving almost always results in an incomplete capture of the selected web content, the integrity of the content is only guaranteed after the archived content has been ingested. Checksums are used to guarantee the integrity of the data within the archival management system. Once the archived content has been ingested into the system, it is also important to keep a trace of the actions that were taken to preserve the content in question in the preservation metadata.[15] It is clear that different kinds of metadata (descriptive, technical, preservation and administrative) have an important role to play in web archiving.[16] The focus of this paper will be on descriptive metadata in web archives.

As web archiving is a relatively new activity, few standards exist. Metadata practices vary wildly between different web archiving initiatives. The British Library for example works with different levels of description for administrative and descriptive metadata. Not all metadata elements are generated for every selected website but possible elements are: rights data and licensing, crawl schedule, title, subject, keywords, short abstract and allocation to a special collection.[17] The National Library of Switzerland on the other hand encodes the following descriptive metadata elements: URL, title, organisation, producer (name, place, canton, country, contact person for granting of rights), language, Dewey classification, keywords and frequency of harvesting.[18]

---

[13] ISO copyright office, *International standard ISO 15489-1 Information and documentation - records management - Part 1: concepts and principles*, 2016, p. 4.

[14] J. Niu, "An overview of web archiving", in *D-Lib Magazine 18*, 3-4, 2012, doi:10.1045/march2012-niu1; L. Duranti, *Requirements for assessing and maintaining the authenticity of electronic records*, 2002, p. 1.

[15] Council of State Archivists, 2019. Retrieved online https://www.statearchivists.org/electronic-records/serp-framework/preservation-metadata/. Last accessed on 09/04/2019.

[16] J. Dooley and K. Bowers, *Descriptive metadata for web archiving: recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group*, 2018. Retrieved from https://www.oclc.org/content/research/publications/2018/oclcresearch-descriptive-metadata/recommendations.html. Last accessed on 28/03/2019.

[17] I. Cooke. (personal communication, February 8, 2018).

[18] Swiss National Library, *Merkblatt Erschließen*, 2017. Retrieved from https://www.nb.admin.ch/dam/snl/en/dokumente/e-helvetica/normen_und_regelwerke/webarchiv_schweizmerkblatterschliessenversion185januar2015.pdf.download.pdf/webarchiv_schweizmerkblatterschliessenversion185januar2015ingerm.pdf. Last accessed on 28/03/2019.

Practices not only differ between national libraries. Traditionally, archives and libraries describe resources in a different manner. In general, cataloguing within libraries is done only at the title level and titles are copied verbatim. Archives on the other hand work with multi-level descriptions of collections for which titles are usually devised.[19] This does not facilitate interoperability of metadata. Recently, efforts have been made to link metadata stemming from different web archive collections. In the Netherlands, for example, eleven institutions have contributed to a national web archiving register.[20] They make their metadata available on a shared platform where the public can use search filters such as archiving institution, name, archived website, archiving tool, reason and year or period of archiving.

The aim of the PROMISE project was to ensure this interoperability between descriptive metadata stemming from the State Archives and the Royal Library for Belgian web archive right from the start. The recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group were adopted.[21] They studied metadata practices related to web archiving at libraries and archives and defined a set of generic descriptive metadata elements. One additional advantage of the OCLC metadata set is that mappings to EAD and MARC21 are included. The vision with regard to access to the Belgian web archive is that users could access it 1) via the shared access platform and 2) via descriptions in EAD in SAM (State Archives Management System) or via MARC21 records in Syracuse (catalogue of the Royal Library). The OCLC metadata set accommodates both approaches and was therefore deemed the right choice. Table 1 shows the 14 descriptive metadata elements as defined by the OCLC. All except one metadata element can be used to describe a specific web archive collection as well as a single website or web page although the content will be different. The title of the website or web page will be transcribed for example, whereas the title of a web archive collection will be devised by the State Archives or the Royal Library.

It is important to mention that not all captured web content will be described in the same manner. The Royal Library and the State Archive will only create detailed descriptive metadata for resources that are part of specific crawls, meaning curated collections of websites based on specific themes (websites of public institutions, national newspapers, music, Belgian comic books, …). For web content that was captured in a broad crawl, meaning a crawl that captures a randomly selected part of the Belgian web, only minimal descriptive data will be provided as the sheer

---

[19] J. Dooley and K. Bowers, *op. cit.* ; J. Niu, *op.cit.*, p. 18. doi:10.1045/march2012-niu1.

[20] Netwerk Digitaal Erfgoed, *Nationaal Register Webarchieven,* 2019. Retrieved from https://www.registerwebarchieven.nl/. Last accessed on 28/03/2019.

[21] J. Dooley and K. Bowers, *op. cit.*; A. Antracoli., K. Bowers and J. Dooley, *Best Practices for Descriptive Metadata for Web Archiving Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group*, 2017. Retrieved from https://www2.archivists.org/sites/all/files/2017WebArchiving.pdf. Last accessed on 09/04/2019.

number of archived websites would make it impossible to describe them in a detailed manner.

Collector	Language
Contributor	Relation
Creator	Rights
Date	Source of description
Description	Subject
Extent	Title
Genre/Form	URL

Table 1: Overview of the descriptive metadata elements[22]

For technical metadata, no such standardisation effort exists. However, contrary to descriptive metadata for web archiving resources that are usually generated manually, technical metadata are created automatically at the time of crawling. A strong link therefore exists between the technical metadata and the WARC file format. Within the PROMISE project most of the technical metadata that are preserved are part of the WARC file either as optional or mandatory fields such as: file format, date and time of capture, file size, target URI, WARC record ID, WARC type, software used, chosen robots.txt policy, http response headers and server IP address. Further analysis is required to determine whether other additional technical metadata elements should be preserved in order to facilitate research within web archives.

The WARC file format

The WARC (Web ARChive) file format is the predominant format in web archiving and is an ISO standard (ISO 28500).[23] It is a container format in the sense that it combines all digital resources that constitute a website and its related information into an archival file. The underlying principle is that all interactions between a browser (web client) and a website (a web server) are recorded in a WARC file. One WARC file consists of multiple WARC records that in turn consist of a record header and a record content block. The content blocks of a WARC file can contain all kinds of formats such as HTML, audiovisual files, images and all other formats found on websites.

[22] J. Dooley and K. Bowers, *op. cit.*
[23] ISO. 2017. *ISO 28500 Information and documentation - WARC file format.*

There are eight different WARC record types: warcinfo, response, resource, request, metadata, revisit, conversion and continuation. Named fields (mandatory or optional) offer information about specific records. For each record type a number of relevant fields has been defined. Figure 2 shows a graphical representation of a WARC file and of a WARC record.

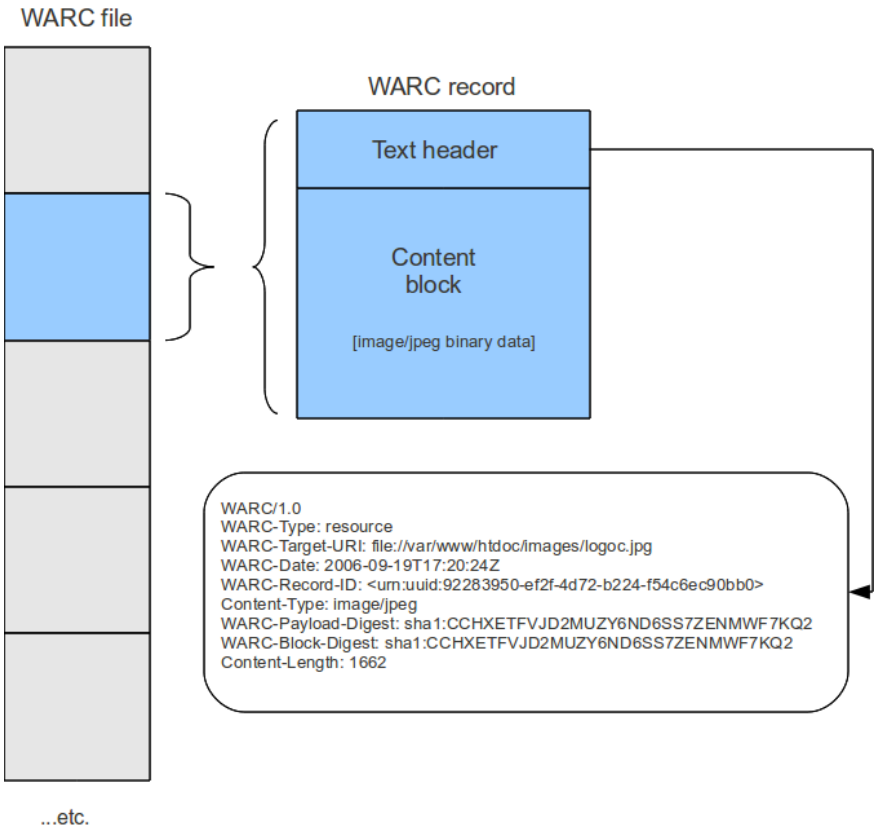

Figure 2: Visual representation of a WARC record inside of a WARC file[24]

Experimenting with adding metadata to WARC records

The reason for experimenting with adding metadata to WARC records was to determine whether it is possible to add metadata directly to the WARC file. In this way, the WARC file is self-sufficient and the risk of information loss is smaller than if the descriptive metadata are stored in a file separate from the WARC file. The web content that was selected for capture within the project was described in a spreadsheet. This approach allowed to get started quickly while ensuring that the

[24] Archivematica, *WARCdiagram*, 2013. Retrieved from https://wiki.archivematica.org/File:WARCdiagram.png. Last accessed on 10/04/2019.

metadata can be migrated later on in the project. Each resource to be archived (i.e., URL) was therefore described using the OCLC metadata set.

It was then explored how to integrate this descriptive metadata into the WARC file. As explained above, the WARC file format is based on a sequence of records of different types, where each record is described by specific named fields. In our case, the most appropriate type of record was "metadata". It contains information that further describes and contextualises captured content that is not mentioned in other record types. Here follows an example of a 'metadata' record.

```
WARC/1.0
WARC-Type: metadata
WARC-Target-URI: http://www.feria-andalucia.com/
WARC-Date: 2019-03-12T14:58:31Z
WARC-Concurrent-To:
<urn:uuid:933898d0-0dcb-4ae8-b820-f3bb56db7bc3>
WARC-Record-ID: <urn:uuid:fa8b6953-c196-42e2-8005-009c467c1c1b>
Content-Type: application/warc-fields
Content-Length: 372

seed:
fetchTimeMs: 7
charsetForLinkExtraction: UTF-8
usingCharsetInHTML: UTF-8
outlink: http://www.feria-andalucia.com/favicon.ico I =INFERRED_MISC
outlink: http://www.feria-andalucia.com/IndexFERIA/Feria-Andalucia.gif E
img/@src
outlink: http://www.feria-andalucia.com/acc.html L a/@href
outlink:
http://www.feria-andalucia.com/IndexFERIA/logoFR.gif E img/@src
```

This type of record allows to add content describing another WARC record of any type, even possibly another "metadata" record. In the experiment, a description of a website is added, so the "response" record corresponding to the URL of that website is described. The named field "WARC-Refers-To" forms the link to this record. This field contains the URN (Uniform Resource Name) of the described record. This URN is actually a UUID (Universal Unique Identifier) generated by the crawler during capture.

Based on this information, a prototype was built that allows the integration of the metadata contained in the spreadsheet into the WARC files. The prototype was developed in Python and the "warcio" module. The latter is used by the

Webrecord pywb toolkit, a complete framework for replaying WARC files. The prototype functions as follows: 1) for each URL in the spreadsheet the tool will identify the corresponding "resource" record in order to obtain its UUID; 2) the tool will then create a new "metadata" record in addition to the already existing metadata record while creating a link to the record corresponding to the URL via the "WARC-Refers-To" named field; 3) it will then add the different metadata in the WARC record content block of the new "metadata" record; 4) and finally save the new "metadata" record in the WARC file.

The same format as for the 'warcinfo' block was used to store the metadata in the content block of the new "metadata" record, namely 'application/warc-fields'. In practice this means that the metadata is presented line by line, according to the structure '<field name>': '<field value>'. An example of a description using this method can be found below:

url: https://statbel.fgov.be
title: Statbel # La Belgique en chiffres # België in cijfers # Belgien in Zahlen # Belgium in figures
creator: SPF Economie, P.M.E., Classes moyennes et Energie # Service Public Fédéral Economie, P.M.E., Classes moyennes et Energie # FOD Economie, K.M.O., Middenstand en Energie # Federale Overheidsdienst Economie, K.M.O., Middenstand en Energie # FÖD Wirtschaft, K.M.B., Mittelstand und Energie # Föderaler Öffentlicher Dienst Wirtschaft, K.M.B., Mittelstand und Energie # FPS Economy, S.M.E.s, Self-employed and Energy
contributor:
language: fre # dut # ger # eng
collector: Archives de l'Etat en Belgique # Rijksarchief in België # Belgisches Staatsarchiv # State Archives of Belgium
genre/form: Site web organisationnel. Website van organisatie.
relation: Archive web services publics. Webarchief overheidsdiensten.
subject: I1_4 # National/Fédéral. Économie # Nationaal/Federaal. Economie # National/Föderal. Wirtschaft
source of description: Description based on contents viewed on August 21, 2018
extent: 1 archived website
rights: In copyright # Freely accessible online
note:

As can be seen in the example above, some fields are multivalued and contain separators between the different values. While this is not a problem in itself, it was deemed important not to rely on local conventions to store metadata. Therefore, it was decided to use the JSON format. In concrete terms, the format for the WARC

record becomes 'application/json' and the content block now contains the following data:


```
{
  "url": "https://statbel.fgov.be",
  "title": [ “Statbel”, “La Belgique en chiffres”, “België in cijfers”, “Belgien in Zahlen”, “Belgium in figures” ],
  "creator": [“SPF Economie, P.M.E., Classes moyennes et Energie”, “Service Public Fédéral Economie, P.M.E., Classes moyennes et Energie”, “FOD Economie, K.M.O., Middenstand en Energie”, “Federale Overheidsdienst Economie, K.M.O., Middenstand en Energie”, “FÖD Wirtschaft, K.M.B., Mittelstand und Energie”, “Föderaler Öffentlicher Dienst Wirtschaft, K.M.B., Mittelstand und Energie”, “FPS Economy, S.M.E.s, Self-employed and Energy” ],
   "contributor": "",
   "language": ["fre", "dut", "ger", "eng"],
  "collector": ["Archives de l'Etat en Belgique", "Rijksarchief in België", "Belgisches Staatsarchiv", "State Archives of Belgium"],
   "genre/form": ["Site web organisationnel", "Website van organisatie"],
   "relation": ["Archive web services publics", "Webarchief overheidsdiensten"],
 "subject": [“I1_4”, “National/Fédéral. Économie”, “Nationaal/Federaal. Economie”, “National/Föderal. Wirtschaft”],
  "source of description": "Description based on contents viewed on August 21, 2018",
  "extent": "1 archived website",
  "rights": ["In copyright", "Freely accessible online"],
  "note": ""
}
```


Based on this experience, it can be concluded that it is possible to include metadata associated with a specific URL in a WARC file. This experiment was based on the OCLC metadata set, but it is also possible to use different metadata sets such as MARC21, UNIMARC, MODS, EAD, METS, Dublin Core, etc. Similarly, the JSON format was chosen to represent the data, but XML, YAML or any other data serialization format can be used. It is even possible to store the same metadata in JSON, XML and YAML in different WARC metadata records.

For this prototype a specific choice was made that is important to mention. The description in the spreadsheet refers to an entire website, but this notion does not exist in the WARC file. Indeed, a website consists of several web pages that each have their own URL. The common point between these pages is an identical

domain name. The choice made for this experiment was to link the metadata with the record that is associated with the domain name. In this way, the duplication of metadata for each page of the described website was avoided.

Conclusion

Web archiving is a relatively new activity for heritage institutions, which explains why standardisation is still mostly lacking. Within the PROMISE project that aims to develop a sustainable web archiving service for Belgium, the State Archives and the Royal Library are striving to ensure as much metadata interoperability as possible by adopting a shared model for descriptive metadata that is based on a study by the OCLC. In addition to that, a mapping to EAD and MARC21 will allow the institutions to also include descriptions of the archived online content into their respective management systems.

The Heritrix crawler is used within the project to capture the selected web content, and the content is stored in the WARC file format, which is the predominant file format in the field of web archiving. Different types of metadata are important in web archiving, but this paper focused on descriptive metadata. The descriptive metadata were created manually using spreadsheets. A prototype was developed that allowed to create a new metadata record in the existing WARC file in order to integrate the descriptive OCLC metadata in the WARC record. This approach diminishes the risk of information loss as the descriptive metadata are stored in the same file as the captured content and not in a separate file. The descriptive metadata are included in JSON in order to avoid relying on local conventions with regard to the use of separators for multivalued fields, but other data serialisation formats can also be used. Other metadata sets than OCLC descriptive metadata can also be integrated in the same manner.

This article focused on one specific aspect of web archiving. As web archiving is a complex activity, other elements are also studied within the PROMISE project, such as the selection and preservation of web content within the specific institutional context of the State Archives and the Royal Library, the surrounding legal framework on legal deposit, copyright and the law on archives, providing access to the web archive etc. By the end of the project (December 2019), all these aspects will be addressed the recommendations for a sustainable web archiving service at federal level in Belgium.