Data Management of Web Archive Research Data

Bolette Jurik, PhD., The Royal Danish Library, baj@kb.dk
Eld Zierau, PhD., IT consultant at the Royal Danish Library, elzi@kb.dk

ABSTRACT

This paper will provide recommendations to overcome various challenges for data management of web materials. The recommendations are based on results from two independent Danish research projects with different requirements to data management: The first project focuses on high precision on a par with traditional references for analogue material and with web materials found in different web archives. The second project focuses on large corpora (collections) of archived web references as basis for analysis.

INTRODUCTION

The focus of this paper is on the part of data management where research data (in form of web material corpora) must be well defined to enable other researchers to verify the research results. The importance of web archives as a research resource is growing and persistent and precise referencing to web material is needed as part of defining the data used in the research [1,2,3,4]. In this paper, we will provide recommendations on how to define web archive corpora in a structured, persistent and managerial way, considering technical issues like size and organizational issues like legal framework.

Web corpus definition

A web corpus is a set of web parts, which have been harvested and preserved in a web archive. A simple example is the definition of parts in a web page. This example will be used for clarification throughout the paper. The example web page shown below has one part containing the html for the page and another part, which is an image on the page.

Figure 1: Example of a webpage with an image component

On the online web, a URI is used as a locater (URL) at the current resolving time. In a web archive, there is no specific time attached to the image on the web page, and therefore it is up to an interface

program like the Wayback Machine to search for archived versions of the image and then choose between the different versions [5]. Thus, implicitly the program uses an extraction algorithm to find the collection (corpus) of parts to give access to the full web page.

An alternative way to define the web page is to provide an explicit collection or corpus definition with all the web parts needed to access the web page. In the above example, the reference could point to a corpus containing the references to the parts

Basis for recommendations

The presented suggestion is based on results and experiences from two independent cases that have investigated different requirements to data management of web data. The two cases were conducted by different organisations and with different focus both regarding researchers and computer scientists. The two research projects had different focuses:

- A *context project* (on contexts of literary works) focused on high precision and persistency on references to web archive materials. The literary field is fundamentally changing in the contemporary digital age: Not only with regard to how the authors publish (e-books, digital poetry etc.), but especially in regard to the changing modes of *reception* (social media, blogs etc.). The project investigated how it could be possible to reference these sources from different web archives with precision and persistency on a par with traditional references for analogue material ([6] row 25).

- A *corpora project* (on extracting results from corpora) focused on challenges with large corpora (collections) from a single web archive where archived web references were used as the basis for analysis. The project was conducted in the Danish project *Practical Data Management* for several case studies [7], where one of the cases was the project *Probing a Nation's Web Domain* [8] (here called the *web sphere project*), which aims at mapping the development of the entire Danish web preserved in the Danish national web archive, Netarkivet [9]. The approach is to investigate so-called spheres annually where web elements can only appear in one version per year.

Our recommendations take into account all relevant results from these projects and the aim is to cover as many archive use cases as possible. The fact that the two projects represent different focuses enables us to cover a large variety of issues that may emerge for corpus related data management for web archive materials. The overall results from the projects points in the same direction, which is the reason we have been able to form the general recommendations presented in this paper.

THE CORPUS DEFINITION REQUIREMENTS

The purpose is to establish corpus definitions that enable other researchers to assess reliability and provenance, as well as to retrace and reproduce research steps. The very basis for continued access is to have corpus definitions that are preserved under a sustainable digital preservation program. Such a program must ensure that the corpora definitions remain readable, understandable and usable

according to preservation purposes for as long as needed. To enable digital preservation actions on a defined corpus, the following requirements must be fulfilled:

- There is a precise and persistent global reference to both the corpus definition and to references used in the definition
- The corpus definition and contents must be on a form that can be maintained for future access

Maintaining future access can only happen if the corpus definition itself has the properties that enables preservation actions. In other words, the corpus definition will require a format suitable for long-term preservation, i.e. having characteristics as a preservation format [10,11]:

- standardised
- well-documented
- open
- understandable
- widely used

Besides preservation related requirements there are also requirements to:

- be technically implementable within a realistic budget
- respect legal and ethical framework for the collection

All of these requirements must be taken into account to make a practice for web corpora definition that can ensure sustainable access to the corpora.

FINDINGS FROM THE PROJECTS

This section describes the different findings in the projects that relate directly to the requirements for continued access to web corpora.

The Technical Framework

The main issue related to implementation has been whether the corpus definition should contain the actual web materials or just references to the web materials in web archives. The conclusion from both projects is the same: it is only possible for corpus definitions to contain references to web material.

The conclusion from the *corpora project* related to the fact that the corpora involved are very large. The intention was to create 10 corpora ranging from 5-30 TB in size. Both for analysis and for preservation, size is an issue. Focusing on preservation, it will not be sustainable to have full preservation of the corpus content in several places due to preservation cost.

The conclusion from the *context project* related to the fact that the corpora covered materials from different web archives. Many web archives are under different legal frameworks, which may prevent the web archives to deliver the actual web data. Therefore, there will be cases where it is legally impossible to include web materials from other archives.

The Legal Framework

Especially for the *corpora project*, the focus was on identifying the legal and ethical issues when preparing the data management plan for web material from Netarkivet. The data in Netarkivet can be both sensitive and copyrighted, and individual access permits are required. Even a single URI can be sensitive data[1], and thus the corpus definition itself can be sensitive data requiring individual access permits.

When sharing a single URI or a very small corpus, it is possible to make sure that it does not contain sensitive data. When working with large corpora, it is necessary to consider other possibilities. The project found that the best solution was to preserve the corpus definition, but restrict access to the same extent as the Danish web archive, and only share the non-confidential metadata. This means that a researcher will be able to find the full corpus definition and apply for access.

Another options considered was to anonymise the corpus definition and share both corpus definition and metadata. This option was discarded, since all research based on contents would become non-reproducible, and further work is limited.

A third option considered was to share corpus extraction algorithm and metadata, e.g. complex algorithms for extraction of web spheres. This option was discarded since there are various aspects that can result in a corpus extraction algorithm yielding different corpora at different points in time. The reason is that algorithms can be hard to preserve and web archives expand over time, e.g. by adding new harvests or including material from other web archives in order to fill holes in collections.

Precise and persistent references

Especially for the *context project*, the focus was to ensure that the corpus definition consisted of persistent and precise references to web parts on an international basis, in order to be able to define corpora across web archives and with a level of detail that exclude any possibility of ambiguity.

The *context project* resulted in proposing a new standard Persistent Web Identifier (here called PWID)[3,4,12] for general, global, sustainable, humanly readable and technology agnostic persistent web references. This differs from Memento [13] and Link Decoration [14] by being very precise about which web archive the reference was found and validated, while the others enable specification of an approximation for what can be found from different open web archives. In short, the suggested PWID includes four main elements:

- <1> web archive identification
- reference to resource:
  - <2> identifier (archived URI)
  - <3> archiving timestamp
- <4> precision of what is referenced

[1] An anonymised example of such an URI is http://activist.com/surveillance-helle-thorning-schmidts-personal-number-is-141266-XXXX/

Examples of PWID URIs related to Figure 1 (on the form pwid:<1>:<3>_<4>:<2>):

- Web page defined by e.g. Wayback:
  pwid:netarkivet.dk:2015-12-03_05.04.37Z:page:http://digitalbevaring.dk/hjemmesider/
- Page html
  pwid:netarkivet.dk:2015-12-03_05.04.37Z:part:http://digitalbevaring.dk/hjemmesider/
- Page image
  pwid:netarkivet.dk:2011-08-21_06.52.06Z:part:http://digitalbevaring.dk/uploads/image.png

Most web archives preserve the URI and the archiving timestamp along with the harvested data, therefore any web part in a web archive can be located by specifying the archived URI and archiving timestamp. Thus, a definition containing specification of the web archive along with this information is agnostic to specific web archive implementations of access technology. The fourth element about precision was needed, since the PWID is also meant to be a precise and persistent reference for web material in literature in general, where precision on page level can be acceptable.

Corpus definition contents

Both the *corpora* and *context project* concluded that the corpus had to contain a reference for each included web part. While the *context project* specified parts using PWID URIs to obtain persistency, the *web sphere project* chose the CDXJ format [15], because it can be parsed directly by the Internet Archive Wayback Machine and offers inclusion of additional data.

Additionally, the *corpora project* looked at other data management practices for other metadata. Especially the DataCite Metadata Schema was used [16], which requires Identifier, Creator, Title, Publisher, PublicationYear and ResourceType with possibility to specify e.g. Description, which was the recommendation from the *corpora project* to describe corpus purpose and context. However, it was also acknowledged that context metadata like algorithms and derived data would benefit from being placed elsewhere with reference to the corpus definition.

The context project took a more minimalistic digital preservation approach, only requiring an absolute minimum of additional metadata: Identifier and Archiving timestamp, but allowing other fields like Title etc. The approach in this project was to place most of the purpose and context metadata elsewhere with reference to the corpus definition.

Corpus definition placement

Both research projects recognized that the actual web material relevant for a data management plan, was best preserved by the web archives with preservation obligations like Netarkivet.

Preservation of the corpus definition is another matter. For the two projects, the best solution would be if Netarkivet could offer to preserve the corpus definitions, as the appropriate preservation program and implementation respecting Danish legal framework already exists there along with specialized knowledge of web archive data. An alternative would be to preserve the corpus definitions in a library or a research infrastructure repository fulfilling the same preservation and legal requirements.

THE RECOMMENDATIONS

In the perspective of the requirements for contents and based on the joint findings from the two projects, we will here present and argue for our recommendations on to how to tackle the various challenges related to documentation of web research data focusing on the corpus definition.

It should be noted that we do not make any recommendation on whether additional metadata to a corpus definition should be placed in the definition itself; or elsewhere with reference to the corpus definition.

Corpus definition

Our recommendation is that a corpus definition is defined as a collection of parts with the follwong minimum contents:

- ***identifier*** is required in order for the corpus definition data to be findable and ultimately re-usable. The identifier should only consists of characters accepted in a URI in order to make it preservable in a web archive.
- ***timestamp*** is required to distinguish different versions registered at different times. The timestamp should be a UTC timestamp in URI in order to make it preservable in a web archive.
- ***contents*** is the set of precise and persistent global references to where the corpus parts can be found. We recommend using PWID URI for these references as it fulfils these requirements, as it is independent of web archive and current web archive technology (which the CDX formats are not).

For us, the important characteristic of the definition is the elements and their structure. We do therefore not want to prescribe how it should be specified in practise. In the below example, we use XML for specification, but use of e.g. RDF would be just as valid. The XML specification defines the simple example from Figure 1, and name it *urn:example_corpus_id*:

```
<collection>
 <identifier>urn: example_corpus_id </identifier>
 <timestamp>2017-05-01 12:04:40Z</timestamp>
 <contents>
  <part>
  pwid:netarkivet.dk:2015-12-03_05.04.37Z:part:http://digitalbevaring.dk/hjemmesider/
  </part>
  <part>
  pwid:netarkivet.dk:2011-08-21_06.52.06Z:part:http://digitalbevaring.dk/uploads/image.png
  </part>
 </contents>
</collection>
```

We recognise that there should be additional optional fields in the structure to allow registration of web research metadata, as there may be a need to specify context metadata within the definition. We do not think a rule can be made for whether the context descriptions should be placed within or outside

the corpus definition. There may be reasons for placing all context metadata separately or within the corpus definition, and it is most likely that a mix will be relevant too.

Corpus definition placement

We would like to suggest that web corpora definitions should be placed and preserved in web archives, as they are just another type of web data. As for the Danish projects, national web archives would be able to provide corpus definition preservation, which respects the local legal framework and offers a sustainable preservation program that fulfils the countries legal deposit laws if any. We know of no web archives that offers such a service yet, but our hope is that there will be further work that can enable this in the future.

The alternative for now is therefore to follow the scheme for corpus definitions, and place it in a repository with a digital preservation program and respecting the legal issues there may be for the corpus definitions.

Corpus definition reference

How to provide a precise and persistent global reference to a corpus definition is very much related to where the corpus definition is maintained as part of a preservation program. We would like to suggest referencing them using a corpus PWID. If corpus definitions were placed as web archive data in web archive, such a reference could for the above corpus definition example be pwid:netarkivet.dk:2017-05-01_12.04.40Z:collection:example_corpus_id. However, PWID is not an obvious choice as long as the web archives do not archive these kinds of data. The alternative is to register in a research data repository with appropriate digital preservation program and access restrictions, and then use the persistent identifier scheme provided by the repository, e.g. a DOI.

DISCUSSIONS AND FURTHER WORK

Our recommendations are based on some very different research projects. Although we are confident that this will cover most cases, there may be cases that are not covered and therefore can result in adjustments, if it make sense to include them.

We note that the recommendations focus solely on the minimum required components in a corpus definition. There are therefore a lot of additional work that needs to done before standardised corpus definition services can be implemented in practice:

- extend with possible additional metadata
- describe all related data management aspects
- provide guidelines for preservation of extraction algorithms
- provide guidelines for possibilities of registration of context

To our knowledge there is currently no standard specifically concerned with web corpus metadata, therefore such work will be an important first step. It would also be a huge positive step forward, if web archives could offer corpus definition services, and if more effort was spent on tools to build corpus definitions based on e.g. places visited in browsing.

A further step could then be to investigate corpus definitions based on other corpora, e.g. the union of several corpora or a corpus extended with some parts. If a corpus definition can be regarded as a 'part' of a web archive, then such an extension would be straight forward.

CONCLUSIONS

This paper has presented a way to define web material corpora in a precise, persistent, global way, which can fulfil requirements to have sustainable corpus definitions that enable researchers to:

- assess reliability and provenance
- retrace and reproduce research steps
- enable continued work

The proposed definition scheme is based on two very different research projects which both needed web material corpora as basis for their data documentation, and it will cover a large range of similar cases. Based on results from these projects we have argued why the suggested scheme can respect:

- long-term digital preservation requirements
- technical implementation issues
- legal framework issues

Although there is still work to do, this is a big step in the right direction, and it can certainly be used as basis for further work.

REFERENCES

[1] N. Brügger and N.O. Finnemann, 'The Web and Digital Humanities: Theoretical and Methodological Concerns', *Journal of Broadcasting & Electronic Media*, vol. 57, no. 1, p. 66–80, 2013. doi:10.1080/08838151.2012.761699.

[2] S. Lawrence, D. M. Pennock, G. W. Flake, R. Krovetz, F. M. Coetzee, E. Glover, F. Å. Nielsen, A. Kruger and C. L. Giles, 'Persistence of Web References in Scientific Research', 2001. Web archive ref.: pwid:archive.org:2016-03-05_22.24.00Z:part:http://clgiles.ist.psu.edu/papers/Computer-2001-web-references.pdf

[3] C. Nyvang, T. H. Kromann and E. Zierau, 'Capturing the Web at Large - A Critique of Current Web Referencing Practices', in *Proceedings of the second RESAW conference* 2017, in press.

[4] E. Zierau, C. Nyvang, and T. H. Kromann, 'Persistent Web References – Best Practices and New Suggestions', in *Proceedings of the 13th International Conference on Preservation of Digital Objects (iPres)*, 2016, pp. 237–46. Web archive ref.: pwid:archive.org:2016-10-12_14.15.31Z:page:http://www.ipres2016.ch/frontend/organizers/media/iPRES2016/_PDF/IPR16.Proceedings_4_Web_Broschuere_Link.pdf

[5] *Internet Archive Wayback Machine*. Internet Archive, 2017. Web archive ref.: pwid:archive.org:2017-01-26_19.40.01Z:page:https://archive.org/web/

[6] *Oversigt over uddelte midler fra Kulturministeriets Forskningsudvalg 2015* (translation to english: Overview of granted resources from the research board, Ministry of Culture 2015).

Danish Ministry of Culture, 2015. Web archive ref.: pwid:archive.org:2015-09-27_21.15.59Z:part:http://kum.dk/fileadmin/KUM/Documents/Kulturpolitik/Forskning/KFU._Oversigt_over_uddelte_midler_2015.pdf

[7] *Practical Data Management Project Description*. The Royal Danish Library (Statsbiblioteket), 2016. Web archive ref.: pwid:archive.org:2016-11-21_08.53.33Z:page:https://en.statsbiblioteket.dk/data-management/practical-data-management

[8] *Probing a Nation's Web Domain - the Historical Development of the Danish Web*. NetLab, 2016. Web archive ref.: pwid:netarkivet.dk:2016-12-07_01.26.57Z:page:http://www.netlab.dk/research/projects/probing-a-nations-web-domain-the-historical-development-of-the-danish-web/

[9] *Netarchive.dk*. Netarkivet, 2017. Web archive ref.: pwid:archive.org:2016-11-19_17.43.57Z:page:http://netarkivet.dk/in-english/

[10] *DigitalPreservationCoalition Digital Preservation Handbook - File formats and standards*. DigitalPreservationCoalition, 2017. Web archive ref.: pwid:archive.org:2017-04-05_14.24.10Z:page:http://www.dpconline.org/handbook/technical-solutions-and-tools/file-formats-and-standards

[11] *Sustainability of Digital Formats: Planning for Library of Congress Collections*. Library of Congress, 2016. Web archive ref.: pwid:archive.org:2017-05-09_12.09.16Z:page:https://www.loc.gov/preservation/digital/formats/intro/format_eval_rel.shtml

[12] *Provisional pwid URI scheme registration at IANA*. Internet Assigned Numbers Authority, 2016. Web archive ref.: pwid:archive.org:2016-12-01_11.32.42Z:page:http://www.iana.org/assignments/uri-schemes/prov/pwid

[13] *About the Memento Project*. The Memento project, 2017. Web archive ref.: pwid:archive.org:2017-05-18_06.28.47Z:page:http://mementoweb.org/about/

[14] *Robust Links - Link Decoration*. The Memento project, 2016. Web archive ref.: pwid:archive.org:2016-12-31_15.49.57Z:page:http://robustlinks.mementoweb.org/spec/

[15] *OpenWayback CDXJ File Format 1.0*. International Internet Preservation Consortium, 2017. Web archive ref.: pwid:archive.org:2017-05-08_13.46.52Z:page:http://iipc.github.io/warc-specifications/specifications/cdx-format/openwayback-cdxj/index.html

[16] *Welcome to DataCite*. DataCite, 2017. Web archive ref.: pwid:archive.org:2017-02-05_01.51.05Z:page:https://www.datacite.org/