LARGE-SCALE COLLECTIONS UNDER THE MAGNIFYING GLASS: FORMAT IDENTIFICATION FOR WEB ARCHIVES

Clément Oury
Bibliothèque nationale de France
Web archive Preservation Manager, Digital legal deposit

ABSTRACT

Institutions that perform web crawls in order to gather heritage collections have millions – or even billions – of files encoded in thousands of different formats about which they barely know anything. Many of these heritage institutions are members of the International Internet Preservation Consortium, whose Preservation Working Group decided to address the issues related to format identification in web archive.

Its first goal is to design an overview of the formats to be found in different types of collections (large-, small-scale…) over time. It shows that the web seems to be becoming a more standardized space. A small number of formats – frequently open – cover from 90 to 95% of web archive collections, and we can reasonably hope to find preservation strategies for them.

However, this survey is mainly built on a source – the MIME type of the file sent in the server response – that gives good statistical trends but is not fully reliable for every file. This is the reason why it appears necessary to study how to use, for web archives, identification tools developed for other kinds of digital assets.

1. BACKGROUND

Since many years, heritage institutions recognized the need to keep the memory of the material that public institutions, businesses and individuals produce and distribute thanks to the Internet. In 2003, some of them decided to group together within the International Internet Preservation Consortium (IIPC). The goals of the consortium are to collaboratively build collections of Internet content, to promote web archiving and "to foster the development and use of common tools, techniques and standards for the creation of international archives". The IIPC is currently made up of more than forty institutions. They generally use – possibly along with other techniques – crawling software, or robots, to explore the web and retrieve content that they will hold for the long term. The sets of documents harvested and produced by these robots are called web archives.

At first sight, from the point of view of formats, web archive collections may appear to be a preservation nightmare. There is no need to recall here the huge number of files harvested by crawl engines. Even the most focused archiving project has to tackle millions of files – see the Harvard University Library, whose Web Archive Collection Service dates back only from 2009 and that already has to preserve 14 million files. These figures rise to hundreds of millions of files per year for those performing crawls of entire top level domains (.au, .fr), not to mention the huge collections of Internet Archive, which in less than 15 years of existence has gathered more than 150 billion files.

The second main issue is that virtually all kind of formats are likely to be available on the Internet. At the same time, when a crawler harvests files online, it gets very little information about the formats of the documents it is capturing. The only indication generally available is the MIME type of the file that the server sends to the harvesting robot, in the http response header. Unfortunately, this information is often badly specified, peculiar (we found at the BnF a curious "application/x-something" MIME type), or even totally wrong (for example, a gif image may be indicated as text/html – webmasters do not see it as a problem for rendering, because a browser is able to read gif files directly).

In short, web archiving institutions generally have millions – or even billions – of files encoded in thousands of different formats about which they barely know anything. Heritage institutions tend therefore to turn to identification tools developed in order to ensure the preservation of other kind of digital material – or developed for other purposes than preservation.

This is the reason why the Preservation Working Group of the IIPC (or PWG) acknowledged the need to specifically address this critical issue through a dedicated work package. In this paper, we will present the goals of this work package and its methodology. We will then look at the first outcomes, and finally present future work.

2. RELATED WORKS

Several studies have been done in order to characterize parts of the web, particularly national web domains. Their goal is to analyze the main features of the websites and web files related to a single country: notably the number of domains, the number of files per domain, the number of hyperlinks between websites[1]… In these studies, we generally find a section dedicated to

[1] See for example [2] for the Danish web, [6] for the Australian web or [8] for the Portuguese web. R. Baeza-Yates et al. proposed in 2006 a comparative study of the national web characterization of several countries across the world, at various dates between 1998 and 2005 [3].

formats. However, we have not identified any works specifically dedicated to file format analysis. On the other hand, there are some – even though rare – studies that examine the ability of identification tools to deal with web archives. In 2007, Bart Kiers from the Dutch National Library tested the behaviour of Jhove and Droid on web archives [5]. The test sample was limited to ten small and medium size websites, grouping 40 000 unique objects for a total uncompressed size of 2.2 Gb. Two years latter, Andrew Long from the National Library of Australia tested five format identification tools (Droid, File identifier, Jhove, TrID and the in-house developed tool Lister) on two web archive samples (from 115 000 to 18 million files) [7]. Finally, the Danish National Library and the Aarhus University Library are currently testing the use of Droid and Jhove on a 100 Tb archive [4].

3. OBJECTIVES AND ORGANIZATION

The first objective of the "format identification tools gap analysis" work package was to produce an overview of the main formats generally available in web archives (using the data obtained from a large number of institutions). It is intended to give a brief insight into the formats that were to be found on the web at different times. This is a way to participate in the general PWG goal of describing the "web technical environment" (that is what formats, software, browsers… were used on the web) over time. On the other hand, this overview should help us in comparing different collections, to identify their characteristics and their specificities.

This study is however built on information – MIME types sent in the server response – that is commonly considered unreliable. First, this has been done for practical reasons: this kind of information was the easiest to get from member institutions. Secondly, we made the assumption that even though the information was not reliable for each individual object, it was sufficient, at a larger scale, to reflect the big picture of format distribution. This assumption has been confirmed by the results of the survey. The proportions found for the only institution that used an identification tool (Library and Archives Canada, which directly ran Droid on their web archives[1]) were globally similar, from 2005 to 2009, to those we found for institutions having only sent MIME information[2].

In the survey, a first distinction is made between domain and selective crawls. Domain crawls are launched on a very large number of websites (e.g. 1,7 millions for both the .fr and .au domains in 2010), but the crawling depth is limited. Moreover, domain crawls are only performed once or twice a year. They are generally launched by national libraries in the framework of a law on digital legal deposit. On the other hand, selective crawls are performed on a more limited number of websites (from hundreds to thousands) generally chosen by librarians or archivists. Those websites may be harvested many times a year, and crawling depth is generally better.

Domain crawls are the best way to obtain a representative sample – a snapshot – of the web. According to R. Baeza-Yates *et al.* [3], crawls of national domains provide a good balance between diversity and completeness by including pages that share a common geographical, historical and cultural context but written by diverse authors in different organizations. However, even though data from selective crawls may be considered less representative (since human, subjective selection replaces automatic selection by a robot), they were taken into account because data from selective crawls may be considered as more "valuable" and may thus deserve more costly preservation actions.

It would not have been feasible to gather information for every year; so the survey focuses on arbitrarily chosen years[3]. Finally, we asked people to give only the list of the 50 most ranked formats. The ranking was calculated according to the number of objects in this format. All institutions were indeed not able to compute the number of bytes per format.

So far, we have received answers from ten institutions[4]. We can consider that this sample is representative of the diversity of the different collections IIPC members may hold: three institutions sent data for domain crawls; eight institutions sent data for selective crawls (some institutions sent data for both types of crawls). Finally, Internet Archive sent information on their crawls of the entire web.

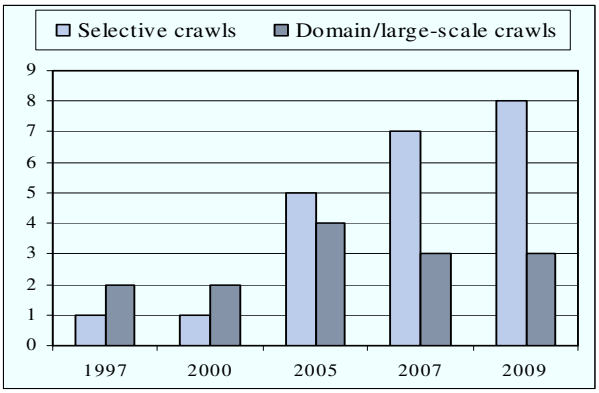

Figure 1. Types of collections in the survey.

[1] Thanks to the Pronom database (http://www.nationalarchives.gov.uk/aboutapps/pronom/), we converted the Pronom identifiers into MIME types.

[2] Note as an exception a surprisingly low number of gif files in the 2005 collection (only 0.8% of the collection against an average percentage of 7%).

[3] It was decided to start from 1997 (date of the first Swedish domain crawl) and to take the years 2000, 2005, 2007 and 2009. We choose to add 2007 because more information was likely to be available for recent years (many institutions didn't start their web archiving program before 2007).

[4] Namely the national libraries of Australia (NLA), France (BnF), Netherlands (KB-NL), Sweden, the Library of Congress (LC), the British Library (BL), Harvard University Library, Library and Archives Canada (LAC), The National Archives of United Kingdom (TNA), and the Internet Archive (IA).

	1997	2000	2005	2007	2009
1	text/html	text/html	text/html	text/html	text/html
2	image/gif	image/gif	image/jpeg	image/jpeg	image/jpeg
3	image/jpeg	image/jpeg	image/gif	image/gif	image/gif
4	text/plain	text/plain	text/plain	application/pdf	application/pdf
5	application/octet-stream	unknown	application/pdf	**image/png**	**image/png**
6	application/zip	application/pdf	no-type/unknown	text/plain	text/plain
7	*application/postscript*	application/octet-stream	**image/png**	**text/css**	**text/css**
8	application/pdf	application/zip	**text/css**	app./x-javascript	app./x-javascript
9	*audio/x-wav*	audio/x-pn-realaudio	application/x-javascript	app./x-shockwave-flash	app./x-shockwave-flash
10	unknown	application/msword	app./x-shockwave-flash	no-type/unknown	**text/xml**
11	application/msword	*application/postscript*	application/octet-stream	**text/xml**	no-type/unknown
12	image/tiff	**image/png**	application/msword	**application/xml**	**application/xml**
13	application/x-tar	**text/css**	**text/xml**	application/msword	application/octet-stream
14	*video/quicktime*	audio/midi	application/zip	app./octet-stream	application/msword
15	audio/x-aiff	*audio/x-wav*	application/x-tar	image/pjpeg	application/rss+xml
16	application/rtf	application/x-tar	image/pjpeg	audio/mpeg	text/javascript
17	video/mpeg	application/x-tex	*application/postscript*	application/zip	image/pjpeg
18	app./vnd.ms-powerpoint	audio/x-pn-realaudio-plugin	audio/x-pn-realaudio	text/javascript	audio/mpeg
19	audio/x-mpeg	audio/x-midi	audio/mpeg	application/rss	application/javascript
20	Javascript	audio/x-sidtune	application/x-gzip	image/bmp	application/atom+xml
21	app./x-shockwave-flash	application/mac-binhex40	**application/xml**	image/x-icon	application/zip
22	**image/png**	image/tiff	application/vnd	app./x-zip-compressed	image/bmp
23	**application/sgml**	*video/quicktime*	app./x-zip-compressed	application/atom	app./force-download
24	**text/css**	application/x-gzip	image/bmp	application/vnd	image/x-icon
25	video/x-ms-asf	chemical/x-pdb	text/javascript	*video/quicktime*	app./vnd.ms-excel
26	x-world/x-vrml	audio/basic	image/jpg	audio/x-pn-realaudio	app./x-zip-compressed
27	application/vnd	application/vnd.ms-excel	*video/quicktime*	video/x-ms-wmv	video/x-ms-wmv
28	image/pjpeg	audio/mpeg	audio/prs.sid	*audio/x-wav*	*video/quicktime*
29	application/x-gzip	application/rtf	video/mpeg	*application/postscript*	app./vnd.ms-powerpoint
30	audio/x-midi	video/mpeg	image/tiff	app./force-download	*audio/x-wav*

Table 1. High ranked formats in large-scale collections, from 1997 to 2009 (increasing formats are in bold, decreasing in italic).

4. FIRST OUTCOMES: GENERAL CHARACTERIZATION

4.1. Web (archive) trends, 1997 to 2009

As a first outcome of this study, we can draw a general overview of the main format trends in web archives. We have compiled information from Internet Archive (available from 1997 to 2005) and from domain crawls of Sweden (1997 to 2009), Australia and France[1] (both 2005 to 2009[2]). Note however that there is an unavoidable gap between the web trends and the web archive trends, because file formats that are hardly harvested by crawlers – flash files, rich media... – are under-represented in archives (and heritage institutions' objective is to reduce this gap by improving their harvesting tools).

It is not surprising to see, all over the period, a strong domination of html, jpeg and gif. It is even more impressive if we look at the percentage of files: for the year 2009, 70% of files are encoded in html, 18% in jpeg, 6% in gif. However, this chart allows us to identify the rise and fall of some formats. We may notice the destiny of png (0.006% of web collections in 1997), which now ranks in fifth place (that is… not even 1.2% of available files). Observe also the increasing rank of css and xml files (while its ancestor sgml disappeared).

On the other hand, some formats that were very popular twelve years ago now rank at a very low place. This is the case of postscript (from the 7th to the 45th place), wav audio files, and even quicktime video files. This is another surprising lesson of this overview: even though we know that the web is increasingly becoming a huge video platform, large-scale crawls don't seem able to tackle the video harvesting issue. The number of captured audiovisual files is increasing (as an example, Sweden crawled 1 300 quicktime videos in 1997, 11 000 in 2005 and 25 000 in 2009), but not as fast as the overall growth of our collections – and definitively not as fast as the percentage of audiovisual content on the web. We will see that selective crawls may provide some solutions to this problem.

From a preservation point of view, however, these figures are good news. Standardized formats are gaining ground against proprietary ones (for example jpeg against gif; xml and png are open formats).

4.2. Comparisons between domain crawls

Statistically, significant differences between collections should not appear in such a mass of data. We can expect

[1] Information used from the BnF for the year 2009 actually dates from November/December 2008.

[2] To compile this information, we calculated the average percentage of formats in different web archives instead of using the total number of files of each collection (e.g. if a format represents 30% of collection A, 20% of B and 10% of C, the average percentage is 20%, even though institution A holds three times more data than the two others). This principle has been applied to all computations. We did so to avoid an over-representation of big collections against smaller ones, which would have prevented all comparisons.

web technologies – and formats – to be equally distributed within the various countries. In fact, if we look at the collections issued from the 2009 domain crawls (France, Sweden, Australia), we find exactly the same formats in the list of the ten most ranked[1]. And we find only 36 different formats in the list of the 30 most ranked.

However, older collections do not show such strong similarities. There is a greater variety of MIME types in the list of high ranked formats for the domain crawls of previous years.

	2005	2007	2009
Top 10	12	10	10
Top 20	25	25	22
Top 30	42	39	36

Table 2. Number of different formats in the list of high ranked formats of the three domain crawls collections, 2005 to 2009.

We can thus conclude that as the web becomes more commonly used, national dissimilarities in the use of web technologies tend to fade away.

4.3. Comparing selective and domain crawls

A similar compilation has been made for collections issued from selective crawls. The goal was also to examine if there were significant discrepancies for collections coming from large- and small-scale harvests, and between small-scale collections from different institutions.

Again, there are no obvious differences between collections. If we compare the average distribution of formats in domain crawls with the average distribution in selective crawls, from 2005 to 2009, we notice few variations. However, a more careful analysis shows some interesting features of specific collections. At the end of the list of the 30 most ranked formats for selective crawls, we find many video formats (such as asf, windows media video or flash videos) that do not appear in domain crawls. Focusing only on formats available in large-scale collections would lead us to leave out these files.

It is also possible to identify characteristics that are related to the nature of the collection. As an example, The National Archives of the United Kingdom are entrusted with the harvesting of governmental publications and websites. This is probably the reason why we notice a larger proportion of pdf and desktop application formats[2].

Moreover, this survey allows us to discover formats that are specific to a collection, over time. For example, the proportion of flash video files which the French National Library (BnF) holds in its 2007 and 2009 selective collections is seven times higher than the average. This last case is explained by the fact that BnF launched in 2007 specific crawls of a video broadcasting platform called Dailymotion, the French equivalent of YouTube.

If we only look at major web archive trends, we will not consider Excel spreadsheets, real audio files or flash video files as being formats that deserve a specific preservation strategy. This is why institutions should also look at their own data in order to assess specific preservation needs. We should not forget the preservation operations won't apply to the web itself – they will be designed for the heritage collections derived from the web.

5. FIT FOR PRESERVATION?

Following on from this, are heritage institutions familiar with such file formats? To answer this question, we can look at a report produced by the National Library of Netherlands (KB-NL). The library conducted a survey on the digital documents held by libraries, archives and museums [10]. From the replies of 76 institutions, they drew up a list of 137 different formats, of which 19 were quoted by seven or more respondents.

5 of these 19 formats only do not figure in our top 20 formats of 2009 domain or selective crawls[3]. On the other hand, the distribution is very different. For example, the most cited format in the KB-NL study is tiff (50 occurrences), while it does not even appear in the top 20 lists for web archives. Similarly, gif and html appear only at the 8th and 10th rank (against 1st and 3rd in web archives). We found similar percentages only for jpeg (2nd rank in both studies), pdf (respectively 3rd and 4th rank) and xml (4th and 8th rank).

The case of tiff files – frequently used for digitization – shows that heritage institutions producing digital documents rarely use the same formats as people that commonly publish online. Yet, can we conclude from this that web formats aren't fit for preservation? To have a first answer, let us refer to the list of "Recommended Data Formats for Preservation Purposes" established by the Florida Digital Archive [9].

Formats are classified in three categories: high, medium and low confidence level. Applying these criteria to the average distribution of 2009 selective crawls (only top 20 highest ranked formats), we can conclude that the formats available on the web are not the worst we can imagine from a preservation point of view (see table 3 below). Note that for some formats (such as html or pdf), there is a different level of confidence depending on the format version. Since this kind of information is not available in MIME type reports, we need to look at the response from Library and Archives Canada – and to assume that its sample is representative. Again, using the 2009 figures:

[1] Excluding the "no-type" format.

[2] In 2005 and 2007, twice the percentage of pdf and word files, five times the percentage of excel files.

[3] TIFF, WAV, AVI, MPEG (2) and MDB files are neither in the domain nor the selective crawls list. BMP is only in the domain crawls list. XLS and PPT are only in the selective crawls list.

- xhmtl files (high confidence) represent 11% of the html files (other versions have a medium confidence grade)[1];
- on the other hand, 98% of PDF files only have a "low confidence" grade. As a matter of fact, PDF-A (high confidence) and PDF-X2 and 3 (medium confidence) respectively represent 0.5 and 1.5% of the total.

MIME Type	Average proportion	Confidence level
text/html	67,979%	High or Medium
image/jpeg	11,885%	Medium
image/gif	6,613%	Medium
unknown/no-type	3,440%	n/a
application/pdf	3,256%	High to Low
text/plain	1,286%	High
image/png	1,182%	High
text/css	0,847%	Medium
application/x-javascript	0,551%	Medium
text/xml	0,444%	High
application/x-shockwave-flash	0,326%	Low
application/atom+xml	0,187%	High
application/xml	0,180%	High
application/msword	0,167%	Low
application/octet-stream	0,114%	Medium or Low
text/javascript	0,104%	Medium
application/rss+xml	0,097%	High
audio/mpeg	0,077%	Medium
application/vnd.ms-powerpoint	0,069%	Low
application/vnd.ms-excel	0,061%	Low

Table 3. Average proportion of MIME types in 2009 selective crawls[2].

6. USING FORMAT IDENTIFICATION TOOLS WITH WEB ARCHIVES

Although this survey provides a first insight into the formats of the collections we hold, this is not enough to guarantee their preservation in the long term. First, it only gives statistical trends: at the level of each individual file, the information is not reliable. No migration operation is possible without such knowledge. Secondly, nothing is said about the format version – which stands as an obstacle for emulation strategies, because we won't emulate the same browser, say, for html 2.0 and 4.0. Therefore, whatever preservation strategy is chosen, relevancy of format information remains a critical issue.

This is the reason why the use of identification tools appears as a necessary step towards a better understanding of our collections. By identification tools, we mean all software that "describes" the format of a specific file. It can range from simple format identification to validation, feature extraction or assessment[3]. This definition may include tools such as Droid, Jhove (v1 & 2) or the National Library of New Zealand metadata extraction tool[4].

Previous reports have already outlined several issues that arise when using identification tools for web archives:

- Some major formats are not supported by characterization tools. For example, neither the NLNZ metadata extraction tool nor Jhove 1&2 are currently able to characterize PNG files. There is no Jhove module for MP3, even though it is the most frequent audio format within web archives…
- Files may not be well formed, which is a problem for identification. This is mainly the case for html files that are frequently hand written or modified. KB-NL reported in 2007 that none of the 20 000 processed html files were considered valid or even well-formed [5]. Let us hope that the growing use of xhtml will reduce this risk.
- Scalability and performance probably remain the major issue for web archives. Tools need to be able to process hundreds of millions of files. NLA report evaluates that it would take 42 days for Droid to identify 18 millions files (0.8 Tb) on a single-core machine, whereas up to a billion files can be harvested in few weeks during a domain crawl [7].

7. FUTURE WORK

The objectives of the PWG are now to organize a collaborative review of the main identification tools. We will build upon the format overview to organize the test protocol and to define the test samples. These tests are intended to assess the efficiency of the tools (notably by providing metrics), and report on any difficulties encountered (e.g. with specific file formats, with the management of container formats, or due to the number of files). Recommendations and best practices for using these tools will be proposed.

Finally, we hope to present a set of enhancements for these tools to address specific web archive issues and requirements. Fortunately, the institutions that are leading the development of the major tools generally hold web archives along with other digital collections, and are also IIPC members.

In addition, test outcomes will also help us to enrich the general overview of the formats in web archives. It will also be necessary to find a durable way to store, update and make available this format overview. An Excel spreadsheet was a convenient way to compile information coming from disparate sources. The work done so far can now be used as a test bench to design a real database, where each IIPC member institution could add its own data.

[1] Note that there is a specific MIME type for xhtml documents: application/xhtml+xml. However, this MIME type is very rarely used, and commonly replaced for convenience reasons by text/html. Even W3C recommends doing so. See http://www.w3.org/TR/xhtml-media-types/.

[2] Figures from 2009 domain crawls are not presented as they show very similar trends.

[3] These categories are defined in [1].

[4] Droid: http://sourceforge.net/projects/droid/
Jhove 1: http://hul.harvard.edu/jhove/
Jhove 2: https://confluence.ucop.edu/display/JHOVE2Info/Home
NLNZ metadata extraction tool: http://meta-extractor.sourceforge.net/

8. CONCLUSION

The first outcomes of this study allow us to avoid an overly pessimistic point of view: even though web archives consist of files over which we have no control, it is not impossible to ensure their preservation.

There is indeed much good news: considering the major trends, it looks like the web is becoming a more and more standardized space. Standard and open formats are gaining ground. Moreover, existing differences between "national" webs are tending to disappear. The second reassuring piece of news is that most files are encoded in a very limited number of formats. Having a preservation strategy for the ten highest ranked formats would be sufficient to render from 95 to 98% of the collection[1].

Yet, this shouldn't lead to an overly optimistic vision. The importance or the "value" of a format does not only depend on the number of files in which they are encoded. This is evident if we choose as the unit of reference not the number of object, but the size. In fact, the ten higher ranked formats (in terms of number of files) generally cover only 50 to 80% of the bytes of the collection[2]. Even the 30 most ranked formats cover only from 70 to 95% of the collection size. This is mainly due to the size of audiovisual files, which are commonly 1 000 to 10 000 times bigger than html pages. Video files may be considered by curators or researchers as more valuable – not only because they hold rich content, but also because without them, heritage web archives collections would not be representative of the "living" web. On the other hand, many html files are not "real" content, but were artificially produced by the robot, for example when it tried to extract javascript links.

Finally preservation actions need to be focused as a priority on file formats that risk becoming obsolete – and this is unlikely to be the case for the major web formats, at least in the short term. This is the reason why institutions may choose to focus on formats that they alone hold: and in this case, having an overview of what is available in other archives will be very useful. This is a way for collaborative work – at national or international level – to provide the tools, knowledge and advice to help institutions to define their own preservation objectives.

9. ACKNOWLEDGEMENTS

The author gratefully acknowledges all those who contributed to the survey. He wishes also to thank J. van der Knijff, S. van Bussel and R. Voorburg from the National Library of the Netherlands for their compilation of the existing literature on web archives formats identification.

[1] For domain crawls, from 2005 to 2009, these 10 formats are: html, jpeg, gif, pdf, plain text, png, css, javascript and shockwave-flash.

[2] Size in bytes is not available for all collections. This ratio of 50 to 80% has been computed from LC selective crawls (2005 to 2009), NLA domain crawl (2009), BnF 2009 domain crawl and 2007-2009 selective crawls.

10. REFERENCES

[1] Abrams, S., Owens E. and Cramer, T. "What? So what?': The Next-Generation JHOVE2 Architecture for Format-Aware Characterization", *Proceedings of the Fifth International Conference on Preservation of Digital Objects*, London, Great Britain, 2008.

[2] Andersen, B. "The DK-domain: in words and figures". Netarkivet.dk, Aarhus, Denmark, 2005. Online: http://netarchive.dk/publikationer/DFrevy_english.pdf.

[3] Baeza-Yates R., Castillo C. and Efthimiadis, E., "Characterization of national web domains", *ACM Transactions on Internet Technology (TOIT)*, 2007. Online: http://www.chato.cl/research/.

[4] Jensen, C., Larsen, T., Jurik, B.O., Hansen, T.S., Blekinge, A.A., Frellesen, J.L. and Zierau, E. "Evaluation report of additional tools and strategies". Kongelige Bibliotek, Statsbiblioteket, Aarhus, Copenhaguen, Denmark, 2009. Still unpublished.

[5] Kiers, B. "Web Archiving within the KB and some preliminary results with JHOVE and DROID". Koninklijke Bibliotheek, The Hague, Netherlands, 2007. Online: http://www.kb.nl/hrd/dd/dd_projecten/webarchivering/documenten/IIPC-PWG-Webarchiving-JHove-DROID-test.pdf.

[6] Koerbin, P. "The Australian web domain harvests: a preliminary quantitative analysis of the archive data". National Library of Australia, Canberra, Australia, 2008. Online: http://pandora.nla.gov.au/documents/auscrawls.pdf.

[7] Long, A. "Long-term preservation of web archives – experimenting with emulation and migration methodologies". National Library of Australia, IIPC, 2009. Online: http://www.netpreserve.org/publications/ NLA_2009_IIPC_Report.pdf.

[8] Miranda, J. and Gomes, D. "Trends in Web Characteristics", *Proceedings of the 2009 Latin American Web Congress*, Washington, United States of America, 2009. Online: http://arquivo-web.fccn.pt/about-the-archive/trends-in-web-characteristics.

[9] "Recommended Data Formats for Preservation Purposes in the Florida Digital Archive". Florida Center for Library Automation, United States of America, 2008. Online: http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf.

[10] Van Bussel, S. Ad Houtman, F. "Gap analysis: a survey of PA tool provision", Koninklijke Bibliotheek, The Hague, Netherlands, 2009. Online: http://www.planets-project.eu/docs/reports/PA2D3gapanalysis.pdf.