Format identification for web archives
Large scale collections under the magnifying glass

international internet preservation consortium

Context and issues

The International Internet Preservation Consortium

The goals of the consortium are:

* To enable the collection, preservation and long-term access of a rich body of Internet content from around the world.
* To foster the development and use of common tools, techniques and standards for the creation of international archives.
* To be a strong international advocate for initiatives and legislation that encourage the collection, preservation and access to Internet content.
* To encourage and support libraries, archives, museums and cultural heritage institutions everywhere to address Internet content collecting and preservation

IIPC members

The IIPC groups together about forty institutions…

➔National libraries

➔National archives

➔University libraries

➔Heritage foundations

➔R&D companies

… from America, Europe and Asia

http://www.netpreserve.org

The Preservation Working Group

The Preservation Working Group (PWG) focuses on policy, practices and resources in support of preserving the content and accessibility of web archives. The PWG aims to understand and report on how approaches used for other kinds of digital resources might be used with web archives, as well as the special characteristics of web archives that might require new approaches. It will provide recommendations for additions or enhancements to tools, standards, practice guidelines, and possible further studies/research.

PWG working fields:

➔Web archive preservation concept and objectives

➔Metadata: capture, packaging, usability

➔Workflows / ingest of web archives in digital repositories

➔Preservation strategies for long-term access

➔Preservation tools gap analysis

➔Web technical environment documentation

➔Organizational issues

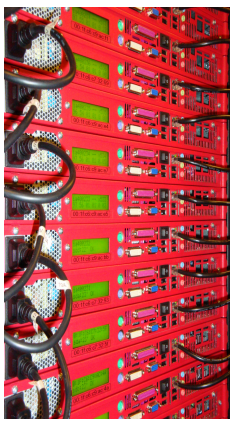
Billions of files in thousands of formats

IIPC members use – sometimes along with other techniques – crawling software, called robots or spiders, to explore the web and retrieve content that they will hold for the long term. From a preservation point of view, these institutions are faced with several important issues:

➔ **Collection size**: this is to be counted in tens of millions of files (smallest and most recent projects), billions (crawls of entire top level domains such as .au or .fr), and even hundred of billions (in the collections of Internet Archive, who over fifteen years has performed worldwide crawls of the web)

➔ **Number of formats**: virtually all kind of formats are likely to be found on the Internet. Most IIPC members are entrusted with the preservation of documents over whose format they have no control

➔ **Insufficient knowledge**: when a crawler harvests files online, the only information it generally gets about the format of the documents is the MIME type of the file that the server sends to the harvesting robot, in the http response header – which frequently turns out to be wrong



File type	Documents	Data
image/jpeg	65806	955 MB
text/html	49653	1.6 GB
image/gif	14979	43 MB
image/png	8660	77 MB
application/x-javascript	2969	47 MB
text/css	2875	50 MB
application/x-shockwave-flash	1682	37 MB
text/dns	1505	155 KB
application/javascript	1118	24 MB
text/plain	1115	7.0 MB
text/xml	954	26 MB
image/jpg	858	5.3 MB
application/xml	643	21 MB
text/javascript	551	11 MB
application/rss+xml	370	9.4 MB
no-type	300	153 KB
image/x-icon	171	659 KB
application/octet-stream	142	2.5 MB
application/atom+xml	75	4.6 MB
application/pdf	41	32 MB
application/rsd	33	28 KB

A MIME type report produced by Heritrix, the harvesting robot developed in the framework of the IIPC

(W)ARC File

(W)ARC Record

Header

Content block

Harvested metadata : date, IP address, MIME type…

HTTP response headers and harvested file: e.g. HTML, GIF, PDF, SWF…

Many IIPC members use the **ARC format** to manage their web archive collections. ARCs are container files where the objects harvested on the web are stored – along with metadata sent by the server or computed by the robot such as harvesting date, IP of the server, MIME response… The **WARC standard** (ISO 28500:2009) is an evolution of the ARC format intended for long-term preservation and access.

Inside web archives

The study: a first attempt to tackle the format challenge

The IIPC Preservation Working Group acknowledged the need to specifically address these issues. Its first objective was to produce an overview of the main formats available in web archives (using data obtained from a large number of institutions). It was intended to give a brief insight into the formats that were to be found on the web at different times. This is part of our goal of describing the "web technical environment" (that is what formats, software, browsers… were used on the web) over time. At the same time, this overview was supposed help us in comparing different collections, to identify their characteristics and their specificities.

Why has it been decided to base the study on information – MIME types sent in the server response – that is commonly considered to be 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.

The study was performed on collections from:

➔British Library

➔Harvard University Library

➔Library and Archives Canada

➔Library of Congress

➔National Library of Australia

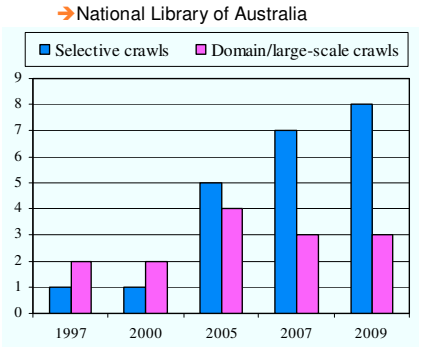
➔National Library of France

➔National Library of the Netherlands

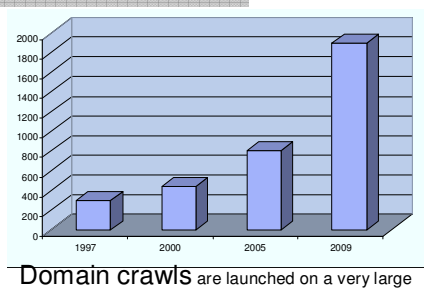
➔National Library of Sweden

➔The Internet Archive

➔The National Archives of United Kingdom



Number of collections for which responses were received

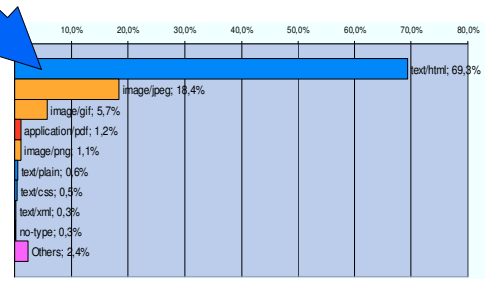


Average number of different MIME types in domain crawls, 1997 to 2009

Domain crawls are launched on a very large number (several millions) of websites, with a limited crawling depth.

They give us a representative sample, a **snapshot of the web**, to identify its major trends – taking into account that some formats (flash files, rich media) are hardly harvested by crawlers

Selective crawls are performed on a more limited number of websites (from hundreds to thousands) generally chosen by librarians or archivists.



Average distribution of formats in web archives for 2009 domain crawls (ranked by number of files)

Average distribution by format types (ranked by number of bytes)

	2005		2009	
Format types	**Domain crawls**	**Selective crawls**	**Domain crawls**	**Selective crawls**
text	39,0%	23,8%	39,9%	30,9%
application	36,2%	44,7%	27,6%	31,2%
image	17,9%	4,7%	18,8%	6,8%
video	4,7%	17,7%	7,2%	23,7%
audio	2,1%	9,0%	6,4%	7,3%
no-type	0,0%	0,0%	0,1%	0,0%

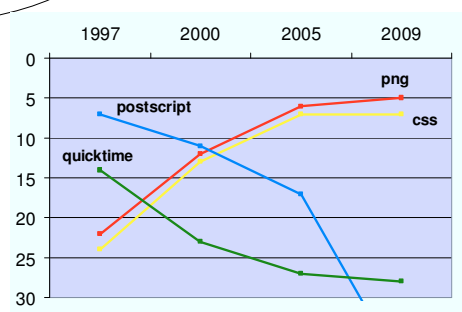
Web (archive) trends

➔ More and more formats…

➔ … but only a few are predominant

➔ Audiovisual files are gaining ground

➔Some are increasing, some are disappearing



Evolution of some formats in web archives (ranked by number of files)

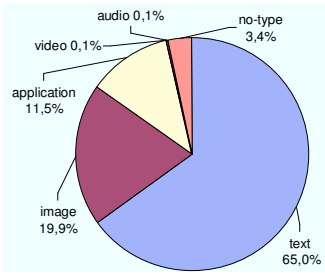
Fit for preservation?

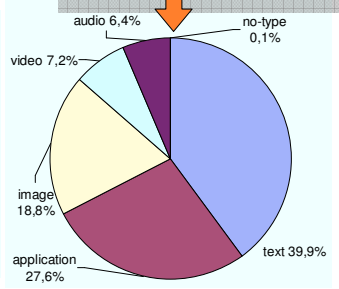
Assessing the risk

According to the "Recommended Data Formats for Preservation Purposes" established by the Florida Digital Archive, formats are classified in three categories: high, medium and low confidence level. Applying these criteria to the average distribution of 2009 domain and selective crawls, we conclude that the formats available on the web are not the worst we can imagine from a preservation point of view. Note that for some formats (such as html or pdf), there is a different level of confidence depending on the format version – and this kind of information is not available in MIME type reports.

Rank	2009 Domain crawls	2009 Selective crawls
1	text/html	text/html
2	image/jpeg	image/jpeg
3	image/gif	image/gif
4	app./pdf	app./pdf
5	image/png	text/plain
6	text/plain	image/png
7	text/css	text/css
8	text/xml	app./x-javascript
9	app./x-javascript	text/xml
10	app./x-shockwave-flash	app./x-shockwave-flash
11	app./msword	app./atom+xml
12	app./xml	app./xml
13	image/pjpeg	app./msword
14	text/javascript	app./octet-stream
15	app./octet-stream	text/javascript
16	app./javascript	app./rss+xml
17	audio/mpeg	audio/mpeg
18	app./rss+xml	app./vnd.ms-powerpoint
19	image/bmp	app./vnd.ms-excel

Confidence grade	High	High or Medium	High to Low
	Medium	Medium or Low	Low





Average distribution by format types for 2009 domain crawls (left: ranked by number of files, right: ranked by number of bytes)

Counting in number of bytes (instead of number of files) changes our perspective on format distribution in web archives. Audiovisual files, that generally hold bigger preservation risks, are more represented. They are also more numerous in collection issued from selective crawls – that is, from websites for which curators ordered specific captures. As these data were most costly to harvest, it make sense to devote more costly preservation strategies to them.

➔ Some good news…

More and more standard formats on the web

Preserving access to ten formats means preserving access to more than 95% of the collection (in number of files)

➔ … some not so good news

Format distribution changes if we look at the number of bytes

Some rare formats may be considered by curators as very valuable

➔ Each institution has to identify the formats it wants to focus on

Using format identification tools for web archives?

Although the MIME type information 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. Secondly, nothing is said about the format version. This is the reason why institutions turn to format identification tools developed for other kinds of digital assets. Previous reports produced by IIPC members have already outlined several issues: many formats – those which are not commonly used by heritage institutions – are not yet supported by these tools; files harvested on the web (especially text files) are neither well-formed nor valid… but the major issue is probably scalability and performance of the tools themselves – they need to be able to quickly process hundreds of millions of files. This is the reason why our goal is now to perform tests, report on the gaps and propose developments for these tools.

% of total	extension	identified as
47%	html	htm / html
25%	php	htm / html / php
13%	jpg	jpg
4%	gif	
3%	jsp	html / xhtml
2%	doc	
2%	xml	
1%	pdf	
1%	png	
1%	txt	

DROID

Positive	Tentative	Not identified
18408		77
9637	14	223
4921		5
1563	4	5
1187		
664	1	
639		1
312	47	
335		
33	139	

% of total	extension	identified as
47%	html	htm / html
25%	php	htm / html / php
13%	jpg	
4%	gif	
3%	jsp	html / xhtml
2%	doc	
2%	xml	
1%	pdf	
1%	png	
1%	txt	

JHove

consistent	well-formed	valid	not well formed
			18485
			9874
	1	4818	107
		1549	23
			1187
not supported by JHove			
	3		637
	25	321	14
not supported by JHove			
		169	3

Reports produced when running Droid (above) and Jhove (below) on a sample of archived websites (source: National Library of Netherlands, 2007)

➔ Test applicability of format identification tools for web archives

➔ Provide metrics and best practices

➔ Report on difficulties and gaps

➔ Recommend enhancements

➔ Propose and fund tool developments

Brief comparison of file format identifiers (source: National Library of Australia, 2009)

Program	JHOVE	DROID	TrID	File Identifier
	Positively Identified (%)			
File Archives	94.95	59.6	91	95
Audio	98.7	4.55	42.85	90
Video	100	67.36	68.75	64.83
HTML	29.74	73.94	46.87	98.16
Microsoft Office	96.08	100	0	98.08
PDF	93.1	100	65.51	100
Raster images	100	93.2	32	64.83
TOTAL SCORE	87.51	71.23	49.57	87.27