

International Internet Preservation Consortium Harvesting Practices Report

Author: Michaela Mayr (Michaela.Mayr@onb.ac.at)
Web@rchive Austria, Austrian National Library

Date of issue: June 10, 2011

Status: Version 2.0

Table of Contents

1	Document Control	3
2	About the Survey.....	3
3	Responders	4
4	Responses	5
4.1	PART A: Harvesting Project Information	5
4.1.1	Background Information	5
4.1.2	Resources: Time /Manpower /Cost	8
4.1.3	The Web Archives	10
4.2	PART B: Policy and Appraisal	18
4.2.1	Scope of Harvesting – Boundaries, Frequency, Limitations	18
4.2.2	Appraisal, Selection Processes, and Access.....	20
4.3	PART C: Implementation and Operational Processes	24
4.3.1	Harvesting Workflow	24
4.3.2	Uses of Tools.....	30
4.4	PART D: Thematic Harvesting	34
4.4.1	Impact of Thematic Harvesting	34
4.5	PART E: Any Other Information.....	40
4.6	PART F: Annexes	42
4.6.1	Annex A.....	42
4.6.2	Annex B.....	43
4.6.3	Annex C.....	43
4.6.4	Annex D.....	45

1 Document Control

Issue	Date of issue	Comments
1	December 23, 2010	For internal review
2	December 27, 2010	For internal review
3	February 11, 2011	With inputs from Grethe Jacobsen, KB Denmark, and Aaron Binns, Internet Archive
4	May 5, 2011	Included surveys from Internet Archive and CDL, inputs from Kris Carpenter and Helen Hockx-Yu
5	June 10, 2011	Included survey data from INA

2 About the Survey

The IIPC Harvesting Practices Survey was developed by National Library Board Singapore in order to understand, analyze and to collate the current Internet archiving processes and experiences amongst IIPC members. The objective was to encourage and support memory institutions everywhere to address archiving and preservation of web resources by providing a benchmark and giving an overview of current web archiving practices.

While the survey aimed to study the current practices on Whole National Domain Crawling, it also included questions on Thematic Crawling practices, and its impact on Whole National Domain Crawling.

The survey was divided into four sections:

- Part A: Harvesting Project Information
- Part B: Policy and Appraisal
- Part C: Implementation and Operational Processes
- Part D: Thematic Harvesting

The IIPC Harvesting Practices Survey opened to members on April 5, 2010, submissions have been collected by NLB Singapore from April to June, 2010.

Due to organizational changes at NLB Singapore the Austrian National Library took on the responsibility for summarizing the survey results in November 2010. The present report, published for IIPC members, intends to give an overview of current web archiving practices. However, it does not represent a manual for the implementation of National Domain Crawls.

Where meaningful, the results have been aggregated. Some answers have been given in detail to serve as an original reference for interested institutions.

The author would like to thank NLB Singapore for developing a comprehensive survey, the reviewers for their valuable input and most of all the participating institutions for sharing their experiences and committing their time and effort.

3 Responders

Of 38 IIPC members in 2010, 17 responded to the survey, representing each IIPC membership region:

Institution	Abbr.¹	Region
The National Library of Israel	IL	Asia
National Diet Library, Japan	JP	
National Library of Korea	KR	
Österreichische Nationalbibliothek (Austrian National Library)	AT	Europe
Netarchive.dk (Royal Library/State and University Library, Aarhus)	DK	
Biblioteca Nacional de España (National Library of Spain)	ES	
Bibliothèque nationale de France (National Library of France)	FR	
British Library	GB	
Institut national de l'audiovisuel (INA)	INA	
Koninklijke Bibliotheek (National Library of The Netherlands)	NL	
Nasjonalbiblioteket (The National Library of Norway)	NO	
Kungliga biblioteket (National Library of Sweden)	SE	
National and University Library of Slovenia	SI	
Bibliothèque et Archives Nationales du Québec (BANQ)	CA	North America
California Digital Library	CDL	
Internet Archive	IA	
National Library of New Zealand	NZ	Oceania

¹ Abbreviations are used throughout the report to refer to survey participants.

4 Responses

4.1 PART A: Harvesting Project Information

4.1.1 Background Information

In this segment, questions aimed to gather general background information of institutions, and web archiving teams. We wanted to study the co-relation of team strength and harvesting processes and outcome.

1. Name of your Institution

See section 3 Responders.

2. Do you have a department or team focusing on Internet archiving? Yes / No

Eleven (65%) have a department or team focusing on web archiving (AT, CDL, DK, FR, GB, IA two teams, INA, JP, KR, NO, NZ only for selective harvesting), six (35%) don't (CA, ES, IL, NL, SE, SI).

3. Please provide some information to your Internet archiving team:

Institutions with web archiving team:

Institution	AT	CDL	DK	FR	GB	IA1*	IA2*	INA	JP	KR	NO	max	min	sum	average
Librarians (full-time):									9	1		9	1	10	0.91
Librarians (part-time):			6.5	6							4	6.5	4	16.5	1.50
Curators (full-time):				4	4	1		1				4	1	10	0.91
Curators (part-time):			2							5		5	2	7	0.64
Software Engineers (full-time):	1	2		4	3	1	6	3		1		6	1	21	1.91
Software Engineers (part-time):		0.5	6	0.3		3	3				1	6	0.3	13.8	1.25
Project Managers (full-time):	1	1		1	3		2	1		1		3	1	10	0.91
Project Managers (part-time):		0.1	1	0.2			2					2	0.1	3.3	0.30

*) IA1: Global Web Archiving Team, IA2: Web Archiving Services Team

Institutions without web archiving team:

Institution	CA	ES	IL	NL	SI	max	min	sum	average
Librarians (full-time):			1		0.5	1	0.5	1.5	0.30
Librarians (part-time):	2			2		2	2	4	0.80
Curators (full-time):					0.25	0.25	0.25	0.25	0.05
Curators (part-time):				3		3	3	3	0.60
Software Engineers (full-time):					0.1	0.1	0.1	0.1	0.02
Software Engineers (part-time):	2					2	2	2	0.40
Project Managers (full-time):				1	0.25	1	0.25	1.25	0.25
Project Managers (part-time):	1	1				1	1	2	0.40

4. How many years has your Institution been harvesting the web:

From 5 months to 14 years: on average 6 years.

5. Do you have a Statutory Act in place for Internet archiving: Yes / No

Twelve (71%) have a Statutory Act in place for Internet archiving (AT, DK, ES, FR, IL, JP, KR, NO, NZ, SE, SI), five (29%) don't (CA, CDL, GB, IA, NL).

6. If possible, please share (as Annex A) with us the current Act your country has in place: Yes / No

Ten institutions have shared their Acts or guidelines. Please see Annex A.

Please rate Q7 and Q8 - with 1 being the MOST and 5 being the LEAST:

7. The most difficult and issues encountered during the course of harvesting are:

Number of responses	1	2	3	4	5	total score
Implementation Processes	4	3	3	3	1	36
Data Management and Maintenance	2	5	3	3	1	38
Implementation Tools	2	3	6	2	1	39
Policies	2	2	2	5	3	47
Scope	5	1	1	1	7	49

8. We would like to resolve and improve the below stated in the following priority:

Number of responses	1	2	3	4	5	total score
Implementation Processes	6	3	4	0	1	29
Implementation Tools	3	3	5	3	0	36
Data Management and Maintenance	0	6	3	4	1	42
Policies	1	1	4	6	2	49
Scope	3	0	2	1	8	53

4.1.2 Resources: Time /Manpower /Cost

This segment gathered findings on the project time management; operational costs and processes; and manpower allocation. The aim was to study the basic available resources needed for a cycle of Whole National Domain harvesting.

9. Which part of the harvesting processes requires the most manpower, time and cost (multiple answers accepted per institution)?

	Number of mentions
Selection of Sites (Scope)	6
Development of tools	6
Monitoring	4
QA of results	3
Problematic sites, technical issues	2
Preparation and planning	2
Data Management and Maintenance	2
Policies	2
Preservation	1

10. Do you plan to expand the team for web harvesting? In which area?

Eight institutions plan to expand their team (additional curators for site selection, additional technical staff/software engineers, additional resources for documenting and workflow processes), nine do not plan to expand (three of them would like to expand but cannot due to financial restrictions).

11. Which area do you wish to further streamline to reduce cost?

Automation is the most important strategy to reduce costs. The main area of automation is the development of tools for quality assurance. Further cost reduction could be achieved by deduplication, improved procedures for preparation, planning, monitoring, site selection, indexing/access and storage/infrastructure (e.g. virtual machines).

12. If possible, please provide the estimated cost for each Whole National Domain crawl cycle.

Three institutions provided figures, ranging from 15,000 EUR for hardware plus approx. 400 hours manpower to 150,000 EUR for inhouse crawls, to 200,000 EUR for outsourced crawls.

In addition, more detailed information was provided by IA:

They perform dozens of national domain harvests at scale, each with very different criteria and resource requirements. In general, costs are measured in weekly intervals, i.e. what resources are deployed and for how many weeks. Hard \$\$ outlays are always included in terms of human resources, hardware for crawling and storage, as well as power to operate and cool equipment, and bandwidth required to collect data (uncompressed). They also deduplicate post download so the raw data collected usually greatly exceeds the compressed data stored. They do not trust the server to report duplicates accurately.

There is also an accounting for opportunity costs that are amortized over the lifetime of a given resource. This helps defray miscellaneous maintenance costs such as disk replacements, software management and upgrades, etc.

As a general rule of thumb they target a capture of 40mil URLs/node/week at peak rate. This rate is hard to sustain for broad and focused crawls after the first 12 hours of collection. The more distributed the crawl, the more efficient the capture. The more concentrated and the deeper a resource, the longer it takes to complete capture. Average through put slows substantially once the smaller sites complete capture. Depending upon the crawl thresholds set, e.g. 1000 URLs per site/5000 URLs per site, etc. the longer the tail.

13. Are there any plans or studies to implement ways to save cost? Yes / No

Seven institutions have plans or studies to implement ways to save cost, nine don't.

14. Do plan to use Cloud Computing? Yes / No

IA regularly uses cloud computing offered by Amazon S3/EC2. They also have begun experimentation with Sun/Oracle Cloud computing. Two institutions plan to use cloud computing (GB uses a small amount as trial already, KR), 14 don't have any particular plans.

15. Please share with us your experience and take-aways from the implementation process.

Following experiences or recommendations have been listed (See Annex for further sources):

- Clarify your scope and goals as early as possible to avoid critical gaps between resources, expectations and outcomes.
- Setting up policies, seed lists, scope etc are the most effort.
- It was difficult to define metadata elements.
- Communication with stake holders is time consuming but very important.
- It is recommended to obtain the seeds from national agencies by formal agreements.
- Develop skills among staff.
- Analyze harvests, it is important to know your domain.
- Challenge to address both development and scale production maintenance.
- Web archiving addresses a moving target.

- Still much unrealized potential, but that has to be balanced with keeping a stable production environment that scales to ever-increasing use.

4.1.3 The Web Archives

The questions in this segment aimed to gain understanding of the web archives - their users and users' behaviour; site features; maintenance; as well as user experience.

16. Please describe the demographic profile of your web archives users (i.e. age, gender etc)

	Number of institutions
(13 institutions responded.)	
General public (JP: mainly for use of congress members)	6
Archive not or not yet public	5
Researchers (DK: at least with Masters' Degree, FR: accredited users of Research Library, at least 18 years old, NO: on site)	3

FR has approx. 100 visitors per month, mostly female, at Masters' Degree level, working in social sciences.

17. Who are your targeted groups of users?

	Number of institutions
(16 institutions responded.)	
General public	11
Academics & Researchers	10
Researchers exclusively	3
Lawyers and people consulting the archive for legal reasons	1
Congress members	1
Libraries, Archives, Museums, Memory Institutions	1
Web masters and site owners (web site recovery)	1

18. What are the difficulties in reaching out to your targeted users?

	Number of institutions
(11 institutions responded.)	
Restriction to on-site access	5
Legal limitations	4
Lack of public awareness, lack of resources for marketing and outreach	3
No direct contact with researchers	1
User interface not designed for disabled people	1
No user interface yet	1
Absence of bulk APIs	1

19. Does your institution promote your web repository and services regularly? Yes / No

Only institutions which already offer public access (or plan to do so) promote their web archives. Methods used are:

- Library websites
- Conferences
- Presentations for interested groups (webmasters, librarians etc.)
- Workshops
- Social media, e.g. Twitter
- Info screens on site
- Print brochures
- On-demand video demonstrations

20. Do you actively solicit feedback from users? Yes / No

Ten institutions do not solicit feedback from users, three do (CDL, GB, KR: for selective harvests).

21. How do you gather feedback from your web repository users?

The following methods and instruments are used to gather feedback:

- Web based feedback forms, surveys
- Web based assessment sessions of designs in progress and filtering the feedback back into the requirements
- Analysis of usage statistics
- Usage studies (based on focus groups)
- Email, telephone, direct contact
- Contact option leading to a help desk ticketing system. Tickets may be promoted to an "enhancement request" category. These are then reviewed during strategic planning and lead to releases around particular functions

22. What are the difficulties in garnering feedback from users?

Mainly the limited number of users, the lack of public awareness, the lack of approval by institutional management to deploy surveys to users or to collect and retain usage logs for more than 30 days, even anonymized.

23. What are the most common complaints from users?

Following complaints have been listed:

- Gaps and inconsistencies in the archive, pages do not fully render (e.g. links do not work, JavaScript missing or not working, pictures are missing, missing material because of robots.txt, etc.)
- Archive only accessible on site
- No catalogue or list of archived sites
- Resources are hard to find
- "Can't find my own site"
- No full text search

- No data mining tools
- Complaints about copyright or privacy infringement (e.g. Website is accessible, which is not desired by content holder.)
- Confusion between live vs. archived websites
- Data is not current, i.e. only current through August 2008

24. What is the favourite feature(s) of your web archives, as regarded by your users?

Following features have been listed:

- Ability to browse the past
- Ability to recover lost sites when they are accidentally lost from the live web
- Personal support by librarians and web-archiving team
- Online access
- Extra depth provided by thematic collections
- Usability of the services
- Site comparison features

25. What are the most popular web archives topics/categories searched by users?

Following categories have been mentioned:

- Political websites
- Web activity
- Government and official publications
- Literature resources
- Media
- Websites within social and technical sciences field
- Thematic collections, e.g. credit crunch, blogs, water resources in California
- Majority of page look-ups occur in the most recent years archived. Index searches on an address mirror live web navigational queries, i.e. the most popular web sites are most frequently searched on.

26. Based on your experience and user trend, what is the common browsing behavior of your users? Do you observe users returning to visited sites, users browsing through different collections etc?

Generally, it is hard to track the browsing behavior of users, since in most institutions usage statistics of staff and external users are mixed. Following examples have been listed:

- Users tend to return to a set of websites as starting point for their research.
- Users frequently compare archived content with live web. Copy/paste URLs from live web to web archive and vice versa and the possibility to shift between online and archived content is highly appreciated.
- Usage becomes more research oriented, as statistics show an increasing number of consultations that last longer than one hour.
- Limited amount of recurrent visitors, some will return to similar searches.

27. How do you ensure your web archives collection stay relevant to users' demand?

Following methods have been listed:

- Develop thematic or event collections with external researchers
- Hold discussions at regular conferences
- Stay up to date with latest web developments
- Content not necessarily relevant today is equally collected, as future demand is unknown.
- Conduct surveys regularly
- Work with graduate student researchers to uncover researcher demands

28. Do you create a unique domain name for easy access to your web archives? Yes / No Do you find it useful? Please provide URL to your site, and please share on your experience with public nomination. Thank you.

Following URLs have been listed:

- CDL: <http://webarchives.cdlib.org>
- GB: <http://www.webarchive.org.uk>
- IA: <http://www.waybackmachine.org> (<http://www.archive.org>)
- JP: <http://warp.ndl.go.jp>
- KR: <http://www.oasis.or.kr>

29. Any areas of the user interface that you would like to improve on?

Following improvements have been listed:

- Enable temporal navigation between different versions of a website
- Full text search with weighted ranking
- Implement data mining applications and services
- Provide topical faceted search results
- Improvements to annotation and personal storage services (e.g. "My web archive")
- Use language filters (e.g. English/Welsh)
- Move some site analysis features for curators out to the public archives (e.g. "show me all the PDFs published on a particular site since a particular date")
- Provide raw WARC format download, both to allow partner organizations to keep a local copy of their content, and also to provide researchers with access to the service as a data set
- Import of metadata records into other repositories, so that pointers to web-archived content can be found side-by-side with scanned materials

30. Please share with us your web repository management processes.

Following examples have been listed:

- No repository in use, ARC files are kept on a file system (AT, NL)
- ARC files in three copies maintained by a SAM-FS solution. One copy on disk and two copies on two geographically separate tape robots (NO)

- Metsfiles generated by Web Curator Tool are kept, each site with UDC like classification (NL)
- Data storage and backup are outsourced to National Federal Computing Centre (AT), collection hosted by Internet Archive (ES)
- Format will be migrated to WARC (GB) or WARC is already in use (JP)
- No preservation actions yet applied (GB)
- Stages of workflow (FR):
 - Pre-processing
 - Harvesting/QA/patch crawls
 - Indexing
 - Storage and copy for backup
 - Post-processing
- Process summary (IA):
 - Data is collected and written to 1 GB WARC files on the crawling nodes
 - WARC files are drained from the crawlers and ingested into the repository system
 - During ingest, the repository system validates the files, indexes them for inclusion in a Wayback CDX, writes them to the primary node, the primary replica is validated and then replicated to a secondary node in the same physical data center, the secondary is validated. If primary and secondary validate, then the file is deleted from the crawler.
 - Once a WARC is in the repository, regular checksums are performed (at least monthly). If problems are detected, the "healthy replica" is used to create new primary and secondary replicas on a new pair and the prior pair is deleted once the new pair has been validated. The file locations are also updated in the index.
 - A third replica is also made to a second physical data center and/or replicated to an institutional repository not maintained by IA.

31. How often do you roll out activities to promote usage to your web archives?

Following examples have been listed:

- One or two meetings per month with researchers, external librarians, website publishers
- Three times a year half-day workshops
- Annual promotion events or conference presentations
- Very Rarely, every 4-5 years
- Every time we get an opportunity
- Never

32. How do you track usage? What are the tools you use?

Following tools have been listed:

- AWStats (AT, FR)
- Google Analytics (GB, in the future CDL)
- Inhouse development (JP), aggregate usage statistics using internal proprietary scripts, logs are not retained for more than 1 to 30 days (IA)

33. How often do you report usage?

Following intervals have been listed:

- Monthly (FR, GB: with summary and trend analysis in annual report)
- Regularly (IA: unique requests and total requests per minute and hourly/daily/monthly, KR: reports for usage statistics, collection statistics, user demand analysis and others)
- Quarterly (DK: number of researchers and topics reported to Steering Group)
- Yearly (CDL: send storage usage report to partner organizations once a year, 60 days prior to billing for storage. Partner organizations have on-demand access to these reports at any time, and for any date range.)
- On demand (JP)

34. Do you find usage report a helpful assessment to how successful the web archives has been? Yes / No

All but one (JP) institutions with public archives find usages reports helpful or are indifferent (DK: only with broader access to archive). Reports on current usage are useful in demonstrating short term success trends to stakeholders and funding authorities. Nevertheless, current usage is not necessarily an indication of future usage, as web archives target future generations to a great extent.

35. If not, how you would prefer to measure the success-rate?

Following examples have been listed:

- A wider public debate about web archiving involving well-known researchers, politicians and policy-makers.
- A more systematic and comprehensive assessment of the percentage of sites which have disappeared (no longer available online) which are still accessible thanks to the web archives, e.g. Change analysis features allow you to produce a report of the files that are no longer on the live web, but that are in the archive.
- Increase in visitor numbers, increase in average time spent on site, reducing bounce rate, unique visitors and time spent on site would be preferred but cannot be easily captured w/o browser plug-in or login. Privacy considerations prevent large scale collection and analysis of these stats today.
- A combination of user statistics and user feedback through questionnaires. It would be interesting to conduct interviews or install user panels.
- Value of the content (cited in publications, library catalogs, used in other contexts)

36. What other quality criteria do you measure for success in your web archives?

Following criteria have been listed:

- Efficient use of limited resources (staff, IT) to acquire a phenomenal and growing quantity of data. Compare this with the current costs of collecting non-

digital data in traditional activities of the library – web archiving is cheap (relatively speaking).

- Web archives are created for future use, so current use isn't the most important aspect. Are the archives good enough for future researchers? Will the archives answer the questions future researchers have? For us this translates to: is the selection good enough? Is the quality of the harvests good enough? Can the authenticity be preserved?
- Increase in rate of permissions granted to archive & display
- Appropriateness of site selection
- Degree of satisfaction of users
- International benchmarking
- URL instances that are publicly accessible
- URL instances per number of TBs ingested per week
- % coverage of the live web (how robust is/are the snapshot/s assembled for a given calendar year); have we accumulated a research quality sample

37. Please share your definition of a successful Internet archiving project, in the areas of short-term and long-term goal.

General goals could be summarized as fulfilling legal deposit laws or selections policies, collection, preservation and access to websites with high quality standards:

- That we are able to fulfill the goal of the law of legal deposit and the objectives of a National Library, i.e. and collect, preserve and give general access to national online heritage
- Archive all online media in accordance with selection policy in the best quality possible, provide access to users in a way they can find the information they want.
- Content is collected based on a well-defined frequency and desired scope in a timely manner to minimize time skew within a collection.
- On the technical level, when we harvest web pages, we aim towards completeness. Thus we have no limit on file size or on how deep the crawler should look. A successful project should get a complete snapshot of the parts of the web we aim to collect, and as few error situations as possible.
- Content is analyzed and independently evaluated to determine the overall quality of the capture based on well-defined requirements for capture and replay.
- Crawl reports are made accessible to the team for each week of crawling and are updated incrementally throughout the harvest.
- Capture and preservation of sites about to be closed, merged or substantially changed, collections of websites featuring a specific event, collections representative of a publishing format / type – e.g. blogs
- Content is indexed as it is ingested into the web repository, and made accessible via the Wayback Machine internally during the harvest and externally post completion of a harvest at an appropriate interval (usually within 6 months of capture).
- Content may also be indexed for full text search and/or mined for meta data and links based on the scope of the project.

Short term goals could be summarized as building sustainable workflows for appropriate selection and quality of harvests and provide access:

- Build a sustainable and complete internal workflow capable of dealing with any web archiving policy decision (whether selective or bulk harvesting), free from any external dependencies (e.g. external services, private software and licenses).
- Capture and curatorial tools are sufficiently easy and effective to preserve at-risk content on the web.
- It will be successful if we can be sure that all the parts of the national domain have been harvested and access to them is provided through a user friendly interface.
- The selection policy has broad commitment
- The web archive that consists of high quality harvests (complete, authentic) based on the selection policy
- Harvest data of good quality

Long Term goals could be summarized as appropriate selection of sites to meet the users' demand, preservation of websites and access for future generations:

- The web archive consists of a broad selection of sites of interest to the general public.
- It will be successful if we come up to our users expectations, regarding the coverage and depth of our collections.
- The web archive consists of authentic sites (and offers tools) that help researchers and journalists find answers to their research questions.
- Preserve the archived data and make it accessible for future generations.
- Save an acceptable and representative sample of the national online cultural production at reasonable cost.
- Analysis, dissemination and integration tools promote the widest possible use of archived materials.
- Use of data by general public

4.2 PART B: Policy and Appraisal

4.2.1 Scope of Harvesting – Boundaries, Frequency, Limitations

In this segment, the questions were set to gather information on the scopes and limits of Internet archiving of the respondents, how the basis of frequency of crawling is determined, as well as the legal liabilities binding the crawling.

38. Please define the scope of your Whole National Domain harvesting:

15 members responded to this question. Six (CA, GB, INA, JP, KR, NL) are not performing whole national domain harvesting at the moment (GB indicated they had plans to do so in the future). Eight institutions (AT, DK, ES, FR, IL, NO, NZ, SE) generally harvest their national top level domains. In addition, all of them are entitled to include content from other top level domains, defined by various criteria, such as geographical information, server location, target audience, language, publisher or ownership of domain.

IA provided more detailed information about the wide range of national domain harvests they perform for national libraries. Each is unique and is based on the legal mandate of an institution, internal policies, and access to third party data sources, among other factors. Most commonly, their whole domain harvests are driven by one or more of the following inputs:

- Resources located on a single or multiple top level domains (for example Sweden uses .se and .nu. New Zealand uses .nz only).
- Resources registered on other top level domains that are hosted within a country specific IP address range. Most often resources on .org/.net/.com are considered but hosts found on other top level domains may be similarly identified/analyzed
- Resources hand-curated by representatives of an institution, usually identifying resources on top level domains that contain content relevant to a national heritage regardless of where the resource is hosted.
- Redirects that point to resources on domains not considered in scope may be included or excluded. There is no standard best practice for this content
- Robots.txt exclusions are usually respected with the exception of embeds (images, animations, videos, audio files, etc.)
- Linked resources that leave the "domain" scope may be included or excluded including PDFs, videos, html pages, etc. but are usually excluded for all MIME types.

39. What would be considered out of scope in your Whole National Domain harvesting?

Everything not listed in the policies is considered out of scope. In some cases further exclusions apply: password protected content (GB), intranets (NO), private content which is not regarded as "publication" and broadcasters' websites due to a shared legal deposit with another institution (FR), anything excluded by robots.txt, linked resources that leave the "domain" scope are often excluded, redirects are often considered out of scope (IA).

40. What are your benchmarks to determine the boundary of scope for your Whole National Domain harvesting?

Five institutions responded that the boundary of scope is defined by their legal mandate and/or collection policies. In most cases, this means again they are targeting their national top level domains. In addition, IA uses information from prior harvests, zone files, legal deposit guidelines and analysis of live web indexes/Alexa Internet Data services.

41. How much time is needed for each cycle of your Whole National Domain crawling?

The duration of national domain crawls ranges from two weeks to eight months. Two institutions crawl shorter than one month (ES, NZ), three between one and three months (DK, FR, SE), three listed durations from three to 8 months (AT, DK in some cases, NO). IA harvests vary from 1 to 8 weeks in length and may result in 10's of millions of URLs up to 1bil+ URLs captured. This information is difficult to compare due to different domain sizes and crawling practices.

42. Do you crawl foreign-based websites with local content? Yes / No

Seven responded with "yes" (AT, DK, FR, GB, IA, NZ, SE), four with "no" (ES, IL in the future, KR, NO).

43. Do you crawl foreign content residing in local domain? Yes / No

Ten responded with "yes" (AT, DK, ES, FR, GB, IA, IL, NO, NZ, SE), one with "no" (KR).

44. Any encounters with legal liabilities while crawling Whole National Domain?

All ten responding institutions did not have any encounters with legal liabilities. However, there are some issues with the ignorance of robots.txt, with unclear legal definitions and public access, and concerns about defamation law suits that could result in requests to change archived websites.

45. Do you use an inventory list for Whole National Domain crawl? How do you derive the list? Please kindly include a sample of a seed.

Four institutions (AT, DK, FR, SE) receive domain lists from their national domain name registries. The seed lists are usually extended by host information from previous crawls or zone files, hosts identified via GEOIP database lookups (e.g. MaxMind GeoIP or other) and manually selected websites (by staff or public). Alexa or other sources are also used to compile seed lists. NZ maintains blacklists and registers for websites that require special harvesting. Seeds lists often contain tens of thousands to hundreds of thousands or millions of seeds.

46. Please provide the frequency for Whole National Domain crawl in a year:

National domain crawl frequencies range from three per year to every two years:

- 3 per year: ES
- 2 per year: SE
- 1-3 per year: DK (would like to increase to 4)
- 1-2 per year: GB, NO
- 1 per year: FR (would like to increase to 2)
- Every 18 months to 2 years: NZ
- Every 2 years: AT

47. What is the size of your most recent Whole National Domain crawl?

Country	URLs (millions)	Size (TB)
GB	4.3	
NZ (2008)	106	4.6
SE	274	13.3
ES	317	
AT	422	6
NO (2008)	485	
FR (2008)	530	19.3
DK (2009)		24
IA	millions of URLs to 1bil+ URLs	250GBs to tens of TBs

4.2.2 Appraisal, Selection Processes, and Access

The questions in this segment were related to policies, aimed at gaining understanding towards the Internet archiving mission and objectives of the responding institutions; appraisal and selection processes. We also wanted to learn more on the access conditions to the archived materials.

48. Can you share with us your current Internet archiving policy?

13 of 14 responding institutions perform selective harvesting as part of their archiving policies and/or legal deposit (ES not at the moment, maybe in the future). Subjects for selective harvesting include media/newspapers, government/administration, science/academics, technology and medicine, business/economy, society and culture and events of national importance. INA collects sites that relate to broadcast TV and radio. NO is also including e-journals and e-books.

In their policies some institutions refer to legal deposits or comprehensive collection mandates (“national cultural heritage”) (AT, DK, ES, FR, IL, NO, NZ, SE), or to permission based harvesting (CA, GB, JP).

Crawls IA performs for IA include all content published to any domain or host that is not excluded by robots.txt. They ignore robots.txt for all embeds. CDL does not set policy; their curatorial partners determine their own policies. They provide some guidelines for formal collection development:

- http://was.cdlib.org/docs/was_collection_plan_guidelines.pdf
- http://was.cdlib.org/docs/was_collection_plan_form.doc

49. How far do your country’s Statutory Acts support web harvesting?

Nine institutions refer to legal deposit legislations as basis for their web harvesting (AT since 2009, DK since 2005, FR/INA since 2006, GB since 2003 for digital publishing, IL, JP for websites of public institutions, NO, NZ). The Swedish law created an exception from personal privacy regulations.

CDL’s rights management practices are guided by the recommendations of the Section 108 Study Group. Section 108 is the part of copyright law that grants certain exceptions for libraries. These recommendations are not law. Guidelines and study group report:

- http://was.cdlib.org/docs/was_rights_management.pdf
- <http://www.section108.gov/>

50. Are there plans for the policy to be revised? If yes, in which area will the changes likely to be implemented?

Five institutions have currently no plans to revise their policies (AT, CA, GB, JP, NO). For DK, FR (decree pending: includes shared responsibility between FR and INA for harvesting the French web), KR and NZ the legal aspects of providing access would be an area of revision. Discussions about personal rights could affect some institutions’ policies. ES is likely to revise its policy regarding scope after an analysis of previous crawls. IA have recently begun top level “/” page, survey harvests of all hosts on every domain. Robots.txt are ignored for these survey harvests. They take snapshots of each host from the live web when the “/” page is harvested. These survey crawls will be performed annually going forward.

51. Are there differentiated levels (internal/restricted/open) of access to the archived sites?

- Access is restricted to web archiving team: CA, NO
- Without differentiation:
 - DK (not yet differentiated, strategy for the future to enable broader access),
 - ES (on site in the future),
 - FR (restricted to accredited readers in research library, plus library staff, different levels for sensitive content could be introduced in the future),
 - GB (online without restriction),
 - INA
 - NZ (maybe differentiated in the future),
 - SE
- With differentiation:
 - AT (Domain crawls accessible on site plus 20 entitled libraries, selective harvesting distributed to libraries by location of site owner, single concurrent user on site for password protected content, embargo possible for max. one year),
 - CDL (archives kept dark, available only to those with accounts and permission for that archive or available to all),
 - IA (All materials are available to researchers upon request, the public can only access materials not actively excluded by robots.txt on the "live" web and there is a minimum 6 month embargo on all harvested material made available to the public.),
 - IL (permission based harvested content online, rest on premises only),
 - JP (internal and public level)

52. How do you categorize the archived sites in the different access levels?

For most institutions this is currently not applicable. AT categorizes by type of harvesting (domain for all/selective for some libraries – defined in legal deposit legislation) and by password protection of sites. DK uses the concept of "safe sites" without sensitive personal data. JP defines the level of access for each site individually.

53. Please indicate the volume or percentage of archived seeds grouped to each access level respectively.

For most institutions this is currently not applicable. IL provides online access for about 10% of harvested websites, JP for about 90%. The rest can be accessed on site. At CDL 35 of 99 archives are publicly accessible.

54. Are there plans to open the restricted access eventually? Yes / No

Six institutions do not have plans to open the restricted access (AT, ES, IA, JP, KR, SE). Seven have plans to do so (CA, DK, FR wants to negotiate access permissions with

publishers in the medium to long term, GB, IL, NL, NO). At CDL curators may make an archive public at will.

55. What are the difficulties in granting access to the sensitive links?

Seven institutions indicate various legal aspects that would concern access for sensitive data, such as protection of personal data (CDL, DK, FR, IA, NL), illegal content, e.g. child pornography, extremism (FR no censorship, no critical case reported so far, NL) or copyright issues (IA, NL).

56. Besides a track record for usage, are there other long-term benefits your institution gained from providing access?

Providing access, where legally permitted, is regarded as an important function of (national) libraries or other web archiving institutions, it increases value and visibility internally (should be considered as ordinary collections by librarians) and externally, and can be used for benchmarking or forging collaborative relationships with other institutions. Without access there was no mandate for collection or preservation. An active user community could also act as quality control.

4.3 **PART C: Implementation and Operational Processes**

4.3.1 **Harvesting Workflow**

By gathering information on implementation processes, we aimed to study how the workflow could be furthered streamlined and improved, and also to look out for potential project pitfalls.

57. Are there any pre-processing/ filtering / quality checks of seeds before the Whole National Domain crawl?

All seven institutions that are pre-processing seeds before domain harvests analyze previous crawls (AT, CDL, FR, GB, IA, NZ, SE). Some filter out unwanted content such as parking sites, crawler traps or sites with almost zero content. Some merge seed lists from various sources, e.g. domain lists from national domain name registries or zone files, seeds nominated by public or librarians, host reports from previous crawls, GeoIP lookup etc. Test crawls are performed using desired scoping rules and configurations, e.g. to verify that seed URLs resolve to a site. The output is analyzed and adjustments are made based on results of the crawl.

58. What are the tools used for this process?

Six institutions (AT, FR, GB, IA, NZ, SE) listed various means, such as inhouse developed tools or scripts, regular expressions in Heritrix configuration, manual crawl monitoring procedures, zone files, MaxMind GeoIP database, host summary reports generated from previous web harvests, spreadsheets, documents or other basic communication tools.

59. Do you set multiple domains/seeds per job? Why?

Eight institutions are setting or will set multiple domains/seeds per job (AT, CDL, DK, FR, GB, IA, NO, SE). Generally, for efficiency of capture, e.g. to limit the numbers of jobs and to balance the load on the harvesting machines and domain servers. FR has created jobs of 3500 domains for the 2010 domain crawl; the number of seeds per domain depends on the result of pre-processing and filtering (at least 7000 seeds per job since two seeds, one with "www" and one without, have been created for each domain). NetarchiveSuite processes harvests in a configurable number of jobs; 3500 provided a good balance between the needs of monitoring and the progression of the crawl.

60. How do you handle files hosted outside the domain?

Three institutions do not save content outside of their domain (ES, GB, NO), one of them is logging and reviewing the seeds which could be added later. IA, NZ and SE take

content if it is in line with their collection policies. DK only harvests the outside domain with a depth = 0, only the link, no further content. FR harvests embeds (images, PDFs, iFrames etc.). AT harvests the content up to a certain size limit that applies for all domains. KR is reviewing the collected data after the harvest and does either delete or keep valuable content.

61. Do you use any filter for images/files hosted out of specified domain for crawling?

Most don't use any filters for images/files, IA and NZ include them if they are embedded in an in-scope page, KR is determining afterwards whether to delete or keep. NetarchiveSuite (DK) allows setting a maximum number of objects per domain and a maximum number of bytes per domain in an XML harvest template. Linked content may or may not be considered in scope.

62. What are your priorities for setting the predefined criteria for Whole National Domain crawl? By size, source, date published etc.

For some institutions, it is most important to get some content from every host. For others a more complete capture of a subset of the domain is most important and there are combinations of both usually tied to caps on number of URL instances collected and/or TBs written. For AT, DK and FR size is an important criterion. Some institutions perform crawls in two steps with limits of 10MB/100MB (AT) and 10MB/6GB (DK). Apart from that, no priorities are assigned.

63. What is your limitations to each of the below stated during Whole National Domain crawl?

	AT	DK	FR	GB	IA	NO	NZ	SE
Number of documents crawled	Limitations possible with max. number of objects and max. number of bytes per domain. Harvesting in two stages: 10MB/100MB per domain	Limitations possible with max. number of objects and max. number of bytes per domain. Harvesting in two stages: 10MB/6GB per domain	Harvesting in two stages: 1,000 URLs /10,000 URLs per domain In total 530-800M URLs		10's of million to 1 billion		Depending on contract with Internet Archive, 130-140M URLs, with option to expand by 50,000 URLs	
Disk space		Until size limit is reached, complete broad crawl about 24TB, still growing	20-30TB	No limit, expectation of size though (110TB)	250GB up to 30TB		No limit, expectation of size though	
Duration to harvest		3-4 months	8-10 weeks		1-12 weeks	6 months	2 weeks to get 140M URLs	Not more than 6 months
Depth	Max. path depth 20, max. hops 25	Max. path depth 20, max. hops 25		Max. hops 100	1000 pages per host up to 10's of thousands of pages per host		Left to Internet Archive	

	AT	DK	FR	GB	IA	NO	NZ	SE
Audio files		Limits for disk space apply		No streamed files, treated in separate process	File size limit is usually 200MBs by default. A file exclusion report is produced and larger files may or may not be included in a patch crawl.			
Video files		Limits for disk space apply		No streamed files, treated in separate process	File size limit is usually 200MBs by default. A file exclusion report is produced and larger files may or may not be included in a patch crawl.			
Other MIME types		Limits for disk space apply			File size limit is usually 200MBs by default. A file exclusion report is produced and larger files may or may not be included in a patch crawl.			
Other types (please state)		Limits for disk space apply			Streaming content is not captured.	Respect robots.txt		

64. How do you monitor crawls?

Crawls are either outsourced (NZ), or if performed by the institutions themselves (AT, DK, FR, GB, IA, NL, NO, SE), monitored with Heritrix interface, logs and reports and NetarchiveSuite (AT, DK, FR). In addition, AT is using Nagios for technical monitoring, SE developed own scripts indicating the number of URLs per domain and server information. NO is monitoring at least every 4 hours and provides an email address for site owners' complaints.

65. How do you overcome issues such as crawl traps?

All institutions performing in house domain harvests (AT, DK, FR, GB, IA, NL, NO, SE), monitor their crawls, look for patterns and block crawler traps, spam or content farms. Heritrix (e.g. regular expressions), NetarchiveSuite (global traps defined in harvester templates, local traps for single domains) and Web Curator Tool provide features to document crawler traps. This information is reused in later crawls. CDL limits captures of single sites to 36 hours.

66. How do you assess the quality of the harvest in terms of quality, and depth?

AT and NL perform manual quality assurance for selectively harvested sites. DK reviews known problematic sites to see how much is harvested. Internet Archive does initial QA for NZ, they also run a small sample of manual QA themselves. FR generates statistics from Heritrix logs and reports, and visually checks a sample of sites. During crawls statistics are reviewed regarding number of URLs collected, number of jobs completed and size of data collected. Post crawl statistics include distribution of top level domains and distribution of MIME types. GB is developing an automated QA tool as module of Web Curator Tool which will spot, report and attempt to restore missing content files. For IA criteria are completeness of capture (usually emphasizing missing files) and volume of unwanted material collected that should be repressed from access and not collected in subsequent harvests. QA reports at CDL include low number of files, error status and time limits reached.

67. How do you index the data post-crawl?

Following methods/tools have been listed:

- Wayback (AT, DK: within Heritrix: crawlstate, linkscope, frontierscheduler, contentsize; later indexing with Wayback viewer, FR: CDX indexes created for access using Wayback Machine, no full-text indexing, IA: always index for Wayback access, sometimes also index for full text search)
- WARC Indexer from Internet Archive (GB, NZ)
- Verity indexer (NL)
- NutchWAX in production, SOLR on development servers (CDL)
- No access, therefore no indexing (NO)

68. How do you metatag your files? Please share your metatagging processes.

Following metatagging processes have been listed:

- Metatags created for ARC record (FR)
- Metatags created for ARC file (FR), crawler output is automatically added to WARC files by default (IA)
- Metatags created for job (AT, DK: name ARC files per jobs and generate metadata files that match the job with metadata in form of harvest setup, reports, logs etc., FR: include all configuration logs and report files generated by Heritrix)
- Name and UDC like classification code added. METS files generated by Web Curator Tool (NL)
- METS profile developed for ingest of web archived material (in WARC format) to corporate digital repository (GB)
- Curators supply subjects and tags (CDL)

69. Do you have in place regular data management and maintenance procedure?

Following procedures have been listed:

- Monthly bit preservation assurance by comparison of files (anything missing?), running of checksum comparisons on the two copies (DK)
- No systematic, automated procedure, but maintenance procedures for hardware/storage checks and replacement (FR)
- Storage and backup outsourced, external institution is in charge (AT)
- Currently no procedures in place (NZ: but intend to develop a preservation plan, GB: want to develop means of identifying and managing corrupt files, virus content etc.)

70. If possible, please provide your configuration file.

Sample from Heritrix broadcrawl settings file (DK) – see Annex C

71. Please share with us your experience in the operational process.

Following recommendations have been given:

- Good preparation and collaboration between curators and crawl operators is a critical key for success, before, during and after the domain crawl. We have put in place weekly meetings and shared documentation/procedures.
- Install an operations manager as dispatcher and first line of problem solving for daily tasks. There are daily considerations and corrections of machine-setup, optimization, firewall changing, disk space adding, harvester processes that unexpected stops or fails.
- It is very important to set up a procedure to manage risks (e.g. complaints from webmasters...) in case something goes wrong during operations.
- Have in place standardized and scalable procedures, which repose on the usage of monitoring applications, and on virtualization of servers to optimize resource usage.

- Streamline operations; one challenge is prioritizing different aspects of the service (crawlers should always be available, ingest service ok to pause for a time).
- Other staff involved in web archiving: operation technicians, software developers and librarians
- Importance of personal experience
- A limited number of shared indicators/statistics, along with a formalised set of goals fixed for the domain crawl before harvesting, help the monitoring operations.
- Use of NetarchiveSuite as an administrative application for setting up harvests, adding seeds and crawler trap filters, monitoring processes, access harvested data, bit preservation and QA.
- Performing test crawls before production crawls helps limit problems during operations.
- Main inconvenience in operational process is the impossibility to foresee/estimate the duration of a job.

4.3.2 Uses of Tools

This segment aimed to understand and study how the harvesting workflow could be optimised with the best uses of available tools, and what tools are needed for future developments.

72. What is the hardware you have in place for Internet archiving?

Following hardware is in use:

- Harvesting outsourced to Internet Archive (ES, NZ)
- JP: mainly IBM servers and storage
- INA: a dual quad core server with 8 TB of disk space
- CA: 4 VM servers and SAN disk space
- GB: 5 dedicated servers, will have 200 Mbit dedicated internet link
- SE: 5 machines for harvesting, archive on a sun-solaris server with attached tape archive, a dedicated machine for development.
- NO: several multiprocessor servers, running a total of eight instances of Heritrix. ARC files stored in 3 copies in a SAM-FS file system (1 copy on disk, 2 other copies on 2 separate tape systems)
- AT: physical machines: 1 admin, 1 indexing, 8 webcrawling, 2 crawl testing (Processors: 2 Intel(R) Xeon(R) CPU E5405 @ 2.00GHz, quad-core, RAM: 2 GB, local hard disk: 500 GB, smb-attached volume: 3 TB (shared by all machines) network interface: 1 Gbit/s operating system: Linux Fedora 8/10/12, Ubuntu 8) Possible to change machines purpose (indexing or harvesting).
- DK Bitarchive storage servers at the State and University Library: 2 Dell PowerEdge 2850 and 2950 with processors : 2 x Intel Xeon 2.8 GHz and Intel Xeon 2.0 GHz both hyperthreaded, RAM: 4GB, local hard disk: 73GB mirrored local 32 TB in SAN (raid 5 and raid 6) and 73GB mirrored local 73 TB in SAN (raid 5 and raid 6), network interface, operating system: Linux red hat(RHEL)

- DK Bitarchive storage servers at The Royal Library: 12 HP DL360 G5 with processors : 2 x QC CPU 2 GHZ, RAM: 3 GB, Controllers: Internal P400, External p800, Storage: 2 x MSA60 one with 3 x RAID 5 with (3 TB) and the other with 2 x RAID 5 (3 TB) , 1 x RAID 5 (2TB) and 1 TB without RAID for temp data, local hard disk: 2 x 72 GB RAID 1 for OS/Software
- FR: 6 physical machines : CPU: 2 x 2,33 GHz Intel Xeon quad-core, 16 GB RAM, 74 GB local disk space, + 15 TB shared disk space for processing, Network : 2 x 1 Gbit/s, Virtualization is used to have multiple crawlers, indexers and pilot machines (2 pilots, 38 crawlers, 5 indexers)
- IA: two generations of crawlers deployed: 5 year old Operton machines with 4-8GBs RAM, a single, dual core processor, 4x1TB hard drives, 2-3 year old red boxes with 4-8GBs RAM, a single, dual core processor and 4x 1TB hard drives, additionally they use Amazon Web Services EC2 instances to crawl.

73. Are there any plans to purchase? Yes / No

Six institutions plan investments such as more storage, increased memory and additional hardware when needed. IA is planning to move to new hardware in the next 6-12 months but will likely assemble their own vs. buy a commercial solution.

74. What is the rate of new hardware requirements (new purchase)?

Detailed information from FR: first architecture put in place late 2006, replaced late 2008, current architecture since beginning of 2010. Estimate that hardware and architecture require partial or complete revision every 3-4 years.

IA uses the same crawlers for 3-5 years. They do not recommend using hardware older than 5 years of age. GB reviews hardware requirement annually. Intervals vary in DK.

75. Is your technical service centre at a different location from the archiving team? Yes / No If yes, how far away?

Six institutions operate with (slightly) different locations:

- Yes (CA: 5km away, CDL: data center is two blocks away and is hosted by the University of California Office of the President. Moving forward, storage services may be located at UC San Diego. IA: within 45 miles)
- In different departments, but in the same building (DK, FR)
- Services outsourced to commercial partners (JP: technical support from software company, 500km away, NL: external server host, AT: storage outsourced to Austrian Federal Computing Centre, same city, content stored in different locations in Austria)

76. Through what kind of channels do you communicate between team members stationed at different location?

	Count
E-Mail	8
Phone	5
Face to face meetings	3

Shared documentation, Wikis	3
Video conferencing	2
Skype	2

77. What are the application(s) your country is currently using to harvest?

Applications in use, solely or combined with Heritrix:

	count
Heritrix	11
Web Curator Tool (incl. Heritrix)	5
NetarchiveSuite (incl. Heritrix)	3
Inhouse development	1
Other	1

78. Is your team researching on other application? Yes / No

	count
Latest developments of Heritrix	4
Have done or will do research	3
Nutch(wax)	2
Wayback	1

79. If yes, what are the objective, direction, timeline, team size and budget for this research?

Future areas of research at FR could include:

- Quality improvement of challenging resource types
- Automated large scale QA
- Sharing crawl information in an internal collaboration database
- Optimization of development plans of NetarchiveSuite and Heritrix

For Wayback research DK is using one developer and a few librarians to test and give feedback. INA has employed an engineer on a 6 month contract to investigate Heritrix and associated IIPC tools. AT used one person month for research of tools in the past. IA is using 4 full time equivalents, 5-10% time over 6 months to migrate operations.

80. Please also provide background of the research team.

More information about research teams (AT, CA, DK, GB, IA, INA):

- Crawl engineer and researcher in web archiving, both with technical background in terms of system administration and programming skills
- Crawl engineers within web archiving team
- Experienced Java developers
- Experienced developer with Masters' Degree
- All crawl operators are university educated software engineers.

81. What application do you use to support the monitoring of the harvest?

See question 77.

82. What applications are you currently using for viewing archived sites?

	count
Wayback OS	10
In house development	3
Viewer in NetarchiveSuite	1
Viewer in Web Curator Tool	1

83. Are there plans to migrate to other applications for viewing? When do you intent to complete the migration?

Two more institutions plan to use Wayback in the future. New solutions might emerge with use of WARC format, e.g. Primo (ExLibris) or WARC tools.

84. What are the difficulties experienced during data migration? Please share with us your experiences.

GB and NZ plan to migrate from ARC to WARC (difficulties in metadata profiling), GB has already performed Pandas to ARC migration in the past (many difficulties compounded by inexperience).

DK experienced that huge amounts of data take very long time to process. E.g. indices for Wayback take months to create and it requires still running machines and applications, which from experience is difficult. Another example: Migration of the total bitarchive (130 TB) from 56 old and outdated PCs to 12 new ones took almost one year.

4.4 PART D: Thematic Harvesting

4.4.1 Impact of Thematic Harvesting

While this survey's main focus was to study the current practices on Whole National Domain Crawling, we also wanted to look into the impact of Thematic Crawling practices on Whole National Domain Crawling. This segment aimed to gain a further understanding on current thematic practices, if any; and to study how it could complement Whole National Domain Crawling, or to complete the cycle of harvesting of web information.

85. Is the Thematic Collection user interface different from the Whole National Domain Collection on your web archive repository? Why?

With respect to access, only NZ distinguishes between thematic and domain crawls. Websites from thematic crawls are available to the public via a catalogue record, whereas domain harvests cannot be accessed at all. At IA there is more metadata available for thematic crawls vs domain harvests.

All other institutions use the same user interfaces (where available), although some have different workflows in place for thematic and domain crawls.

86. Do you have a unique domain to Thematic Collections? Yes / No Why?

NZ provides an URL to a web page explaining the web archive, but websites are only found in the online catalogue. CDL provided the URL <http://webarchives.cdlib.org>. Others do not have a unique domain to thematic collections.

87. Do you intend to expand more intensely on the topic/theme most demanded by users? Yes / No

Eight institutions intend to expand thematic harvesting (AT, FR, IL, INA, KR, NL, NZ, SI), four don't (DK, GB, JP, NO). Not decided yet by IA.

88. Do you accept public nominations to archiving an event or URLs on your web archives? Yes / No

Ten institutions accept public nominations (AT, CDL, DK, FR, GB, IA, KR, NL, NZ, SI), four don't (IL, INA, JP, NO).

89. How do you select thematic seeds? Do you use any tools?

Generally, there are no special tools in use, all have manual processes in place to select thematic seeds mostly with common internet search tools (In addition, NL is investigating the usefulness of automated link frequency tables). In most cases

curators/subject librarians recommend seeds according to selection policies, public nominations are also taken into consideration.

DK reviews 80 websites annually, both for relevance and depth of harvesting. The review is done by collection responsible department heads assisted by an advisory editorial board. FR operates a network of 80 subject librarians and started to work with external librarians and researchers from other organizations. Propositions were managed with an in house tool which will be redeveloped to work with NetarchiveSuite (use excel sheet in the mean time). A special tool for election crawls might be reused for other projects. NL uses a list of 31 topics as base for internet searches. NL and NZ recommend checking webpages for outgoing links to other relevant websites. Portals to discover particular subjects and general awareness (follow recent developments in internet publishing, e.g. internet books, blogs, newspapers, etc.) also help to select relevant seeds. IA uses the Archive-It scope it tool and offline analysis tools for analyzing link structure of prior crawls and crawl reports.

90. What is your limitations to each of the below stated during thematic crawl?

	AT	CDL	DK	FR	GB	IA	INA	NL	NO	NZ	SI
Number of documents crawled		None	Different for each thematic harvest	Defined annually with IT dept., 2010: 200M URLs	None	Usually up to 25mil URLs per harvest but a crawl budget set by the operator/ curator (any size)	None	Default 10,000, sometimes exceptions to 100,000	No specific limits	100,000	50,000 in some cases 100,000 per site
Disk space	Fixed limit, e.g. 100 MB per domain	None	Sum for all thematic crawls 4 TB per year	Defined annually with IT dept., 2010: 20 TB	None	Usually up to 2TB compressed per harvest	None	Not set	No specific limits	1 TB in Digital Asset store plus approx. 600 GB harvesting space	None
Duration to harvest		36 hours (can extend if needed)	Repeated 6 x a day, other types monthly, may only last few hrs	None	None	Up to 3, 5 or 7 days	None	100 hours for most sites	No specific limits	Gov. sites up to 5 days, might stop prior if crawl is slow	None

	AT	CDL	DK	FR	GB	IA	INA	NL	NO	NZ	SI
Depth	Max. path depth 20, max. hops 25	26 hops	Typically 2 levels	None (depending on curators)	None	varied	5	Not set	No specific limits	Complete website if possible	Maximum depth in most cases, some exceptions
Audio files		None	No special, but close monitoring	None	Captured	varied	None	None	No specific limits	Collect if possible	None
Video files		None	No special, but close monitoring	None	Captured	varied	None	None	No specific limits	Collect if possible	Individual (content and file size)
Other MIME types		None	No special, but close monitoring	None	Captured	Flexible, based on curator input via Archive-It tool	None	None	No specific limits	Captured	None
Other types (please state)		None	No special, but close monitoring	No. of seeds limited for each thematic crawl, e.g. regional election 2,000	n/a		None		No specific limits		None

91. Please provide the frequency for thematic crawls in a year:

Frequencies for thematic crawls range from less than once a year to every two hours:

- Less than once a year (IL)
- 1-2 per year (AT, FR, NL, NZ: most annually, political events every three years when elections)
- 3-4 per year (NO: events)
- Daily crawling of news websites (AT starting 2011, FR, NO, DK)
- When necessary/Determined individually for each website (CDL, DK: dependent upon significance of event and resources, GB, KR, SI)
- IA: Nine frequencies including twice daily, daily, weekly, monthly, quarterly, semi-annually, annually, once only, & on demand
- INA: crawl homepages every two hours. Deeper crawls are conducted on a daily and weekly basis.

92. How do you determine the frequency of thematic crawl?

In most cases curators/librarians determine the frequency of thematic crawls individually on a website basis (within constraints of other crawls and technical capacities). Criteria for the definition of the frequencies include:

- Rate of updating of sites
- Risk of disappearance
- Presence of an archive on the site itself
- Relevance of content
- Availability of staff for QA

93. Do you limit the size on thematic crawl? If yes, what is the limit?

Some institution set limits regarding sites per collection (IL: 100), amount of URLs (FR: 100,000 per host, SI: 50,000 per site), storage (AT: 1-2 TB for daily news crawls, JP), depth (INA: 1 to 5 levels from homepage), bandwidth and staff resources. IA sets limits based on document count, time, or data volume. Details see question 90.

94. How often do you check for elimination for Thematic crawls?

Intervals for such checks vary; GB has determined intervals for revisits, NZ checks during each harvest, NL during QA, NO sets a deadline of completion.

95. What is the ratio of time given to Whole National Domain Crawl selection and appraisal to Thematic Crawl selection and appraisal?

For institutions which only perform thematic crawls this cannot be compared, respectively, the effort involved is relative to the objective. In 2009/2010 FR reduced time for thematic crawls due to the first internal domain crawl, but this is not representative for the future. NZ uses approx. the same amount of time.

96. Please provide background to the staff members tasked to the selection of themes and events for Thematic crawls.

Most institutions involve (subject) librarians (CDL, FR: in charge of acquisitions, with specific area of expertise, with university degree from Bachelors' to PhD. Those involved in web archiving attend special training sessions.) or subject matter experts who are fairly non-technical (IA), a trained documentalist with experience in TV & Radio (INA), media researchers, library technicians or members of the web archiving team (NZ: each staff member has a key area of selection, 2 FTE in total for thematic web archiving). A few have installed an editorial board or plan to do so.

4.5 PART E: Any Other Information

97. Please share with us your comments and/or any other experience in your course of work that you deem valuable and helpful to improve the Best Practices in Internet archiving.

FR:

We have learnt a lot by sharing/benchmarking our procedures and plans with other institutions (Internet Archive, Netarchive (Denmark), National Library of Norway, British Library, National Library of New Zealand). We think the present report can help other institutions to have access directly to a compilation of these best practices rather than to ask each institution individually.

We would like to have regular access to configuration files/information for domain crawls run with Heritrix by other institutions, and to review/compare them annually for instance.

It took us five years to acquire knowledge and expertise on how to run a domain crawl, and we are not done yet. Documentation and internal transfer/sharing of expertise is critical because of staff turnover and the need for sustainability of the complete workflow.

We believe Heritrix has been developed for the purposes and the environment of Internet Archive and does not necessarily meet the specific needs and constraints of institutions such as national libraries. These needs should be taken into account to facilitate the implementation and ownership of new versions.

NL:

In our selective approach quality of the harvest is a big issue. We are trying to find an effective and efficient approach to do quality assurance. Currently we keep an administration of all quality issues found in a harvest in the annotation fields of WCT. We do that in a very structured way (using syntax and a classification) so that we can use the results to automate processes. Using this approach we hope to get a better view of things that can go wrong, how to fix them or what issues to give priority. QA is mainly a manual process. In the future, tools might be able to automatically detect quality issues making QA more effective and efficient.

Detecting and preventing crawler traps is an issue we are struggling with. Much improvement by automation seems to be possible.

An issue of concern is the storage space needed. Deduplication might offer a solution. It is investigated how to store harvested websites in the next digital repository. Besides

preservation we need fast and flexible enough access to the web archive without the need for separate storage for access and for preservation.

NZ:

We outsource the work of harvesting to Internet Archive and would recommend this approach. We will eventually bring it in-house.

Key decisions before harvest include:

- Exact scope and how the seeds will be assembled and how this accords with legal deposit.
- Robots policy and other major crawl parameters
- Communications plan: how will you inform webmasters and website owners that the harvest is approaching (assuming you will)

Plan how to respond to media enquiries and other stakeholder feedback during the harvest: a domain harvest is likely to be quite a high-profile activity and one that attracts negative comment in the media and in blogs and social networking sites. You should have a plan for responding to these in place before you begin.

4.6 PART F: Annexes

4.6.1 Annex A

Statutory Acts/Guidelines:

- AT:
http://ris.bka.gv.at/Dokumente/BgblAuth/BGBLA_2009_I_8/BGBLA_2009_I_8.pdf (German)
- CDL: While the U.S. does not have a specific act in place, CDL relies on the Section 108 Study Group recommendations to guide rights recommendations:
<http://www.section108.gov/docs/Sec108StudyGroupReport.pdf>
CDL guidelines based on this report:
http://was.cdlib.org/docs/was_rights_management.pdf
- DK: <http://www.kb.dk/en/kb/service/pligtaflevering-ISSN/lov.html> (English)
- ES:
<http://dglab.cult.gva.es/Legislacion/Orden1971ReglamentoDepositoLegal.pdf> (Spanish)
<http://www.bne.es/opencms/es/LaBNE/Adquisiciones/DepositoLegal/Legislacion/index4.html> (Spanish)
- FR: French Heritage Code (Code du Patrimoine) – see “Titre III” on legal deposit for clauses relating to Web archiving, articles L131-1 to L132-6:
<http://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000006074236&dateTexte=20100420> (French)
Brief summary of the law on legal deposit of the web on BnF website:
http://www.bnf.fr/en/professionals/digital_legal_deposit/a.digital_legal_deposit_web_archiving.html (English)
- GB: http://www.opsi.gov.uk/acts/acts2003/ukpga_20030028_en_1
- IL:

Israel: Copyright Act, 2007

The translation reprinted below is an unofficial translation, prepared by the Israeli Ministry of Justice [...]

30. Permitted Uses in Libraries and Archives

(c) Copying of a work by entities of the type prescribed by the Minister, for purposes of preservation, is permitted; The Minister may prescribe types of works which will be subject to this subsection, conditions for the execution of copying as well as conditions for the grant of public access to copies that were made in accordance with this subsection.

Israel: The National Library Act – 2007

The translation reprinted below is an unofficial and partial translation, prepared by The National Library of Israel. [...]

23. Permitted Uses by The National Library

(4) Copy for the purpose of preservation of Internet web sites or works stored in such sites. Access by the public to copies made according to the present paragraph, will be carried out following the conditions and the restrictions to be established by the Minister

of Justice, and following the agreement of the Minister of Education; such prescriptions will be determined by taking into account, among other factors, their consequences for the copyright owners on such works.

- NO: <http://nb.no/fag/for-utgjevarar-og-trykkeri/pliktavlevering/legal-deposit> (English)
- NZ: National Library of New Zealand (Te Puna Maturanga o Aotearoa) Act 2003: <http://legislation.govt.nz/act/public/2003/0019/latest/DLM191962.html>, National Library Requirement (Electronic Documents) Notice 2006: <http://legislation.govt.nz/regulation/public/2006/0118/latest/DLM381515.html>
- SE: <https://lagen.nu/2002:287> (Swedish)

4.6.2 Annex B

Further resources:

- DK: <http://netarkivet.dk/index-en.php> (English),
Brief recapitulations of our experiences in our newsletters published online:
August 2009:
http://netarkivet.dk/nyheder/Newsletter_Netarchive_dk_august2009.pdf,
August 2008:
http://netarkivet.dk/nyheder/Newsletter_Netarchive_dk_august2008.pdf?highlight=20080908
- FR: <http://iwaw.net/08/IWAW2008-Lasfargues.pdf> (English)

4.6.3 Annex C

Sample from Heritrix broadcrawl settings file (DK):

```
<?xml version="1.0" encoding="UTF-8"?>
<crawl-order xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="heritrix_settings.xsd">
  <meta>
    <name>default_orderxml</name>
    <description>Default Profile</description>
    <operator>Admin</operator>
    <organization/>
    <audience/>
    <date>20080118111217</date>
  </meta>
  <controller>
    <string name="settings-directory">settings</string>
    <string name="disk-path"/>
    <string name="logs-path">logs</string>
    <string name="checkpoints-path">checkpoints</string>
    <string name="state-path">state</string>
    <string name="scratch-path">scratch</string>
    <long name="max-bytes-download">0</long>
    <long name="max-document-download">0</long>
    <long name="max-time-sec">0</long>
    <integer name="max-toe-threads">50</integer>
    <integer name="recorder-out-buffer-bytes">4096</integer>
    <integer name="recorder-in-buffer-bytes">65536</integer>
    <integer name="bdb-cache-percent">40</integer>
    <!-- DecidingScope migrated from DomainScope -->
  </controller>
</crawl-order>
```

```

    <newObject name="scope"
class="org.archive.crawler.deciderules.DecidingScope">
    <boolean name="enabled">true</boolean>
    <string name="seedsfile">seeds.txt</string>
    <boolean name="reread-seeds-on-config">true</boolean>
    <!-- DecideRuleSequence. Multiple DecideRules applied in order with last
non-PASS the resulting decision -->
    <newObject name="decide-rules"
class="org.archive.crawler.deciderules.DecideRuleSequence">

        <map name="rules">
            <newObject name="rejectByDefault"
class="org.archive.crawler.deciderules.RejectDecideRule"/>
            <newObject name="acceptURIFromSeedDomains"
class="dk.netarkivet.harvester.harvesting.OnNSDomainsDecideRule">
                <string name="decision">ACCEPT</string>
                <string name="surts-source-file"/>
                <boolean name="seeds-as-surt-prefixes">true</boolean>
                <string name="surts-dump-file"/>
                <boolean name="also-check-via">false</boolean>
                <boolean name="rebuild-on-reconfig">true</boolean>
            </newObject>
            <newObject name="rejectIfTooManyHops"
class="org.archive.crawler.deciderules.TooManyHopsDecideRule">
                <integer name="max-hops">25</integer>
            </newObject>
            <newObject name="rejectIfPathological"
class="org.archive.crawler.deciderules.PathologicalPathDecideRule">
                <integer name="max-repetitions">3</integer>
            </newObject>
            <newObject name="acceptIfTranscluded"
class="org.archive.crawler.deciderules.TransclusionDecideRule">
                <integer name="max-trans-hops">5</integer>
                <integer name="max-speculative-hops">1</integer>
            </newObject>
            <newObject name="pathdepthfilter"
class="org.archive.crawler.deciderules.TooManyPathSegmentsDecideRule">
                <integer name="max-path-depth">20</integer>
            </newObject>
            <newObject name="acceptIfPrerequisite"
class="org.archive.crawler.deciderules.PrerequisiteAcceptDecideRule">
            </newObject>
            <newObject name="globale_crawlertraps"
class="org.archive.crawler.deciderules.MatchesListRegExpDecideRule">
                <string name="decision">REJECT</string>
                <string name="list-logic">OR</string>
                <stringList name="regexp-list">
                    <string>.*\?cmno=.*year=. *</string>
                </stringList>
            </newObject>
            . . . here follows a very long list of crawlertrapfilters . . .

```

4.6.4 Annex D

FR: Summary of methodology

Step	To do	Action	Actors	Data set
1: Preparing the crawl				
Organizing and managing the crawl				
	Scheduling and planning		Librarian / IT	
	Distributing different tasks between the actors			
Defining the check list				
	Defining target (times and number of URLs)			
	Defining scope		Librarian	
	Choosing the seeds list (AFNIC, curated list, previous crawls...)		Librarian / IT	
	Defining special collections and seeds		Librarian	
	Identifying Junk data		Librarian / IT	?
	Defining settings : politeness, budget, replenish		Librarian / IT	
2: Creating the seed list				
	Testing AFNIC list : checking if domains are online or not		IT	AFNIC list
	The .fr and .re host list coming from the previous broad crawl		IT	CDX, host report
	Curated seed list / 1		Librarian	Curator tool list
	Curated seed list / 2		Librarian	Curator tool list
	Merging and sorting the seed lists into one list		IT	
3: Creating jobs			IT	
	Distributing the various lists on different servers			
4: Performing a test crawl				
	Test crawl		IT	Merged list

5: Running the jobs				
6: Monitoring the crawl				
	Editing and analyzing frontier reports			Frontier report
		Producing metrics	IT	
		Analyzing metrics	Librarian / IT	
	Visual quality control			
		Indexing the data	IT	
		Analyzing the results	Librarian	
	Meetings for feedback		Librarian / IT	
	Modifying parameters and settings, creating overrides		IT	
	Budget management		Librarian / IT	
7: Stopping the crawl				
	Deciding to stop the crawl		Librarian / IT	
	End of indexing process		IT	
8: QA				
	Editing and analyzing general metrics of the crawl		Librarian / IT	
	Editing a manifest of all the generated ARC files		IT	
	Visual QA		Librarian	
	Performing a "monkey" QA			
	Patch crawl			
9: Results and feed back				
	Summarizing the broad crawl		Librarian	
	Building a knowledge base for subsequent crawls		Librarian / IT	

Metrics from Frontier report:

- Total number of queues
- Number of actives queues
- Number of queue which have reached their maximum budget
- Size of the most important queue by server
- List of the queues which contain 100 000 URLs and more