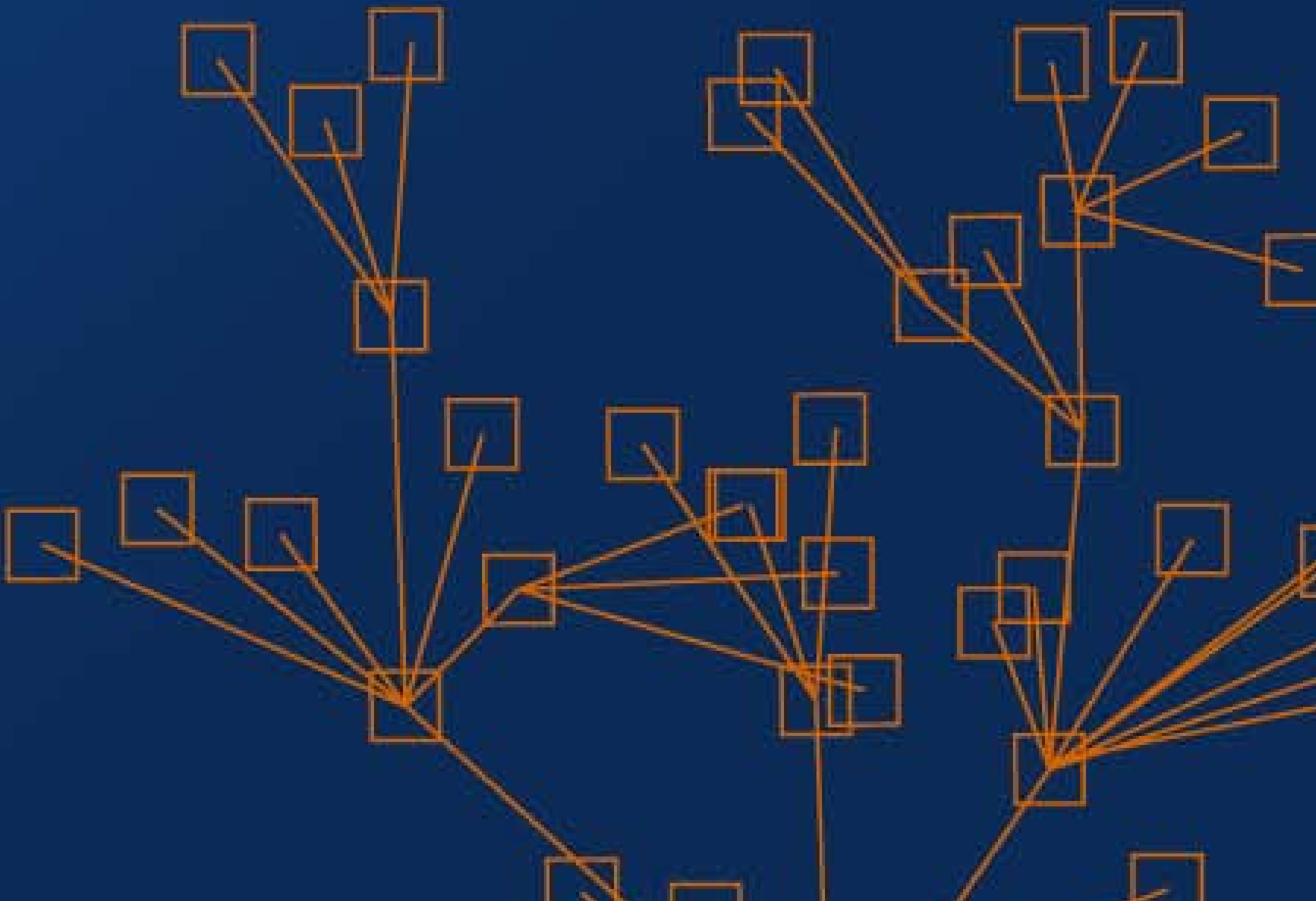
netpreserve.org
international internet preservation consortium

Use cases for Access to Internet Archives

IIPC Access Working Group

May 2006 | Version 1



1 Introduction

Archives of material published on the Internet are still in their infancy; the oldest archive (www.archive.org) is a mere 10 years old. It is conceivable that just as the advent of the Internet has forever changed the way people publish and exchange information, the availability of archives of Internet material will result in new and innovative usages of archives. Much of the current effort related to Internet archives focuses on collecting and preserving the information. The success of the Internet however is based on the easy access to information. It seems reasonable to assume that the ultimate success of any Internet archive will be measured by the means of access the archive provides to the preserved material. In order to gain a better understanding of the expectations and requirements users of Internet archives are likely to have, this report identifies and specifies a number of possible usage scenarios of an Internet archive. Attempts at predicting usage of new technologies are by nature uncertain and likely to miss important applications. The list of usage scenarios presented in this report should thus be considered an opening statement in an ongoing dialogue with relevant user groups.

The web is a medium of its own and a web archive has both the character of a library and an archive, and the collection must be available for research and should be available for other use as well. Therefore it is assumed that the archiving institution will have unrestricted access to the archive. Whatever the use it must not violate the privacy of individuals or the legitimate interests of those who published the material on the public web, and it is not intended to replace use of the public web. Access will very much depend on the legal foundation of each web archive. This varies from country to country and in addition the laws of copyright and personal data may apply as well. Therefore provisions must be made for controlling the access to the documents according to those laws. In addition there are many commercial and "fair use" issues that must be considered when providing access. It is obvious that it will never be accepted that harvested restricted data can be accessed everywhere by all. An example of this is a newspapers article index that requires payment for use. There are plenty of other similarities.

To avoid a massive use of loose terms like "the system" or "the archive", we shall use the name "ArcSys" for the imaginary archive that the example users are accessing.

This document is the result of work and discussions by the members of the IIPC Access Group. The draft was written by Niels H. Christensen of the Royal Library of Denmark.

This report relates to the Prototype report that is based on some of the usage scenarios and attempts to illustrate what interfaces may be needed by developing mock-up prototypes of the functional components suggested by the use cases.

Table of Contents

2 General considerations

In this chapter we will try to identify those dimensions of a given use case that seem to be of special importance when designing an access tool for an Internet archive.

2.1 Where and to whom is a service offered?

A specific facility in ArcSys is not necessarily useful to all users. E.g. linguistic researchers and patent solicitors/lawyers probably have completely different requirements to the user interface of the archive. Therefore potential target groups appear in the scenarios in the next chapter.

Furthermore, there are a number of different "service localities" where ArcSys meets the user and a specific functionality may be offered at several of these.

First of all, ArcSys should have a public web site and this web site should have relevant, updated information about the archive. For example the users should be able to get information about how much material/documentation has been harvested at a given time.

Another "locality" is a ArcSys web site for registered users (that is, users with a login to ArcSys). This locality is distinct from the fully public web site because it is possible to offer personal work tools (e.g. electronic bookmarks). Finally, it is worth noting that registered users may be important to the further development of the ArcSys because they can contribute with useful ideas of improvements and extensions of the system.

A third locality that is relevant to most national libraries is referred to as secure networks. Services offered on a "secure network" are limited to computers, which are placed at the domicile of the archive or at entrusted institutions (e.g. university libraries), which are connected to the archive through secure communication.

Finally, it is possible to offer service in connection with a concrete enquiry. If this is implemented one can forward an enquiry (e.g. by e-mail) to the employees at ArcSys. They will carry out the necessary research and provide a reply.

2.2 Unique facilities

What information can ArcSys deliver that cannot be obtained any other way? We would like to focus on those functionalities that are unique to ArcSys, i.e. those functionalities that require storage of complete URL-objects (unlike indexing as in search engines) and/or use of the time perspective.

2.3 Technical challenges

Some facilities are technically simple while others require a large programming or software integration work in order to be implemented. The needed requirements are of course important factors when considering which services that are to be offered by ArcSys.

2.4 Other Aspects

Motives and incentives

What do we wish to obtain with the functionalities we decide to offer and what do we wish to avoid?

The general reputation of ArcSys will to a considerable extent be controlled by the ways in which different groups use ArcSys. Therefore we should be careful in our selection of the facilities we offer the users of ArcSys. A special issue is how to make it attractive to allow storage of ones documents and materials in the archive.

The question about the reputation of the archive is more political than technical and therefore this issue will not be dealt with in this paper.

The legal setting

What are we permitted to offer within the present and coming legal setting, in particular acts on copyright and processing of personal data.

The legal foundation of the archive’s facilities varies from country to country and we have chosen not to discuss this issue in this paper.

3 Cases relating to the usage of data

3.1 Journalistic documentation

A TV-journalist is preparing an item about the debate on primary schools with focus on a local politician Jane Jones. The journalist recalls that Jane Jones once discussed the topic eagerly on a debate site p-schools.org, but the homepage does not exist any longer. The journalist requests the URL http://www.p-schools.org, and ArcSys returns a list of dates on which this web site has been harvested. By choosing a specific harvest the journalist now is able to see and use the page as it appeared at the time the debate took place.

Comments:

Type of user: Professional.

Technical requirements: Nordic Web Archive (NWA) and The Wayback Machine (WM) offer this facility.

3.2 Digital management

3.2.1 Main case

The municipality has imposed a fine on Jane Jones because she, until November 2003, delivered her garden refuse in plastic sacks and not in paper sacks, which was demanded, from April 2003. However, she has noted on the homepage of the municipality that up to December 2003 it was permitted to deliver the garden refuse in plastic sacks, and therefore she holds that the municipality is not entitled to impose this fine on her. Jane Jones states the address of the homepage of the municipality, specifies the period December 2003 and writes the search string “garden refuse collection”. ArcSys returns the content of the page that she acted according to. Furthermore, the page has a fixed reference in ArcSys. This reference can be enclosed in her complaint over the fine.

Comments:

User Type: Common citizen.

Technical requirements: Uncomplicated. The scenario illustrates use of persistent identifiers.

3.2.2 Alternative case

Jane Jones has a link to the regulations of the municipality, but she is not sure of what she saw and when she saw it. She keys in the link and ArcSys lists the times at which the page has been harvested and whether it has changed in between the harvests. Evidently all harvests from February 2003 until January 2004 are identical. ArcSys does not show the contents of the page but she is informed that the alteration in December 2004 was, that the sentence "Garden refuse for refuse collection must be delivered in paper sacks" has been added to the regulations.

Comments:

Technical requirements: Adequately intelligent processing may be complicated.

Other: See the above. This version may even be more useful.

3.2.3 Alternative case

Jane Jones is officer in charge in Department for Evening Classes under the Ministry of Education. In connection with an excursion a school leader has contacted the Department in order to get the Ministry's approval of the implementation of the excursion. Jane Jones pieces together guidelines related to similar cases from documents on the homepage of the Ministry and the homepage of the municipality of the school and approve the excursion. Later on the municipality complains over the handling of the case, but Jane uses ArcSys (in the same way as in one of the above versions of the scenario) to justify her decision.

Comments:

User Type: Professional.

Other: As the above versions of the scenario.

3.3 Documentation in relation to employment

A lawyer is representing Jane Jones in a case against her former employer ACME. To this end she is looking for documentation for her period of employment at the firm ACME between March and July 2004. The lawyer keys in "Jane Jones" and specifies the period and the domain "acme.com". ArcSys returns a list of URLs which has been stored from "acme.com" and which match "Jane Jones". In each URL the dates of storage is stated. The lawyer may click on the specific URLs and in this way be able to look at the contents. He prints out the pages relevant to the case.

Comments:

User type: Professional.

Technical requirements: Uncomplicated.

3.4 Civil Case Evidence

Jane Jones learns that a competitor of her company copied the appearance, including trademarks, of her business website, and engaged in a mass mailing to customers directing them to the spoofed website, where some were tricked into supplying proprietary information. By the time legal action is contemplated, the competitor has removed the offending material from the website, but copies exist in the ArcSys. Jane Jones keys in the URL and relevant dates of the offending website, and wishes to receive evidentiary-quality printouts of the relevant pages, which may require signed declarations of ArcSys personnel.

Comments:

User type: Litigants, lawyers.

Special requirements: ArcSys staff may be required to testify, in writing or in person, about the reliability of captured data.

3.5 Criminal Case Evidence

A prosecutor learns that the suspect in an arson case may have expressed an intent to destroy the targeted premises on websites that are no longer active. The prosecutor uses the known URLs, or text searches on the suspect's name or aliases, to discover the suspect's published writings before the arson incident. The prosecutor wishes to receive evidentiary-quality printouts of the relevant pages, which may require signed declarations of ArcSys personnel.

Comments:

*User typ*e: Lawyers (prosecutors and defence attorneys), police, defendants.

Special requirements: ArcSys staff may be required to testify, in writing or in person, about the reliability of captured data.

3.6 Inquiry into prior art of patents

An employee at the Patent and Trademark Office has received an application for a patent and is looking for any possible prior art of this patent. The employee keys in a list of key words that relate to the described technology. ArcSys returns a prioritized list of web pages that match the key words. The pages are grouped according to domains. For each page its URL and a summary is given. By clicking on the URLs of the sites the employee can see the pages in the archive.

Comments:

Type of user: Professional.

The facility is also offered at: Unique for ArcSys (search engines can only find prior art which is presently documented on the internet).

Technical requirements: Builds on known technology.

4 Cases relating to user interface

4.1 Basic searching for a known URL or site

Jane Jones wants to check what information was on the homepage of her municipality in 2004.

- She knows the URL and types it in the URL-box of the ArcSys User Interface.
- After clicking the <search> button she gets as a result a list of the different versions of the homepage. The list tells the date when the page was harvested.
- She selects the version she wants to study.
- The page is displayed.

Comments:

Type of user: Common citizen.

Technical requirements: Search UI:

- Direct access by typing in the URL.
- Presentation of different versions according to date.
- Navigational support (e.g. click here to go back to the: search page / hit list / previous version / next version, etc.).

4.2 Advanced Searching for a certain subject

Jane Jones is studying public web sites of different municipalities between 2004-2006 in order to determine how the municipality presents local information to the community. She has narrowed her study to tourism information and local real estate information.

- She locates municipal sites by doing a free text search or by URLs she knows.
- She uses the UI functions to narrow the results down by date.
- She gets a hit list which displays some information about the sites plus the link to the archived page.
- She chooses one hit from the list. She then requests and gets a display of all the versions of that one site.
- She requests one version and the ArcSys displays the page. The links in that page point to the archived versions of each page. So she is able to navigate the archived site as it was on the internet thus being able to find all relevant documents to her study.

Comments:

Type of user: Professional.

Technical requirements: Full text indexing. Search UI:

- All the same as in Use Case 1.
- + Free text search.
- + Narrowed by date/time/format.
- + Hit list, relevance ranking.

Presentation:

- + Navigation between found documents, navigating within the archive (links turned to the archive).
- + Navigation bar and help files for the User.

4.3 Advanced Searching with version comparison, linking information

Jane Jones is studying public web sites of different municipalities between 2004-2006 in order to determine how the municipality presents local information to the community. She has narrowed her study to tourism information and local real estate information. Furthermore she also wants to study how single documents have evolved through time (version control, difference detection). Also one aspect is to study how different documents relate to one another (linking information). The ArcSys should be able to provide some statistical data about these issues.

- She uses the UI functions to narrow the results down by date.
- She also narrows it by to cover only 25 different municipalities (narrow by site/host/domain).
- She performs the query.
- She gets the result list but she wants to change the narrow-down parameters. She clicks the "modify search" button and does the changes.
- She gets the results she wants and she selects one page relevant to her study. She views it and then returns to the result list and selects "show the versions".
- She gets the list for all the versions of that particular page. She selects two of those pages and asks the ArcSys to determine the differences between these two pages.
- ArcSys displays the difference results (percentage, word count, link count, image count).
- She returns to the versions list and selects one version. She then asks the ArcSys to provide linking information for that version of a page.
- ArcSys displays the linking information (known incoming links within the archive, outgoing links, internal links, etc.).

Comments:

Type of user: Professional.

Technical requirements: Search UI:

- All from the above Use Cases 1-2.
- + Narrow down functions (domain, site).

Presentation:

- + Versions can be selected for further analysis.
- + Module to perform and display linking information.
- + Module to perform and display differences between versions.

4.4 Theme collection usage, browsing style usage

Jane Jones is studying public web sites of different municipalities between 2004-2006 in order to determine how the municipality presents local information to the community in the light of communication theories. She has narrowed her study to tourism information and local real estate information. To her amazement the archiving institution is regularly collecting municipal web sites into a specialized collection.

- She selects this collection <select collection>.
- The ArcSys returns a page with further selections.
- She fills in the time period she is interested in.
- She selects host(s) from a drop down menu (Municipal X).
- ArcSys displays the FrontPage.
- She navigates her way to the pages of interest to her.

Comments:

Type of user: Professional.

Technical requirements: Archive must be arranged in a way that supports this type of browsing features.

Other: Requires a special type of archiving policy.

4.5 Personified features

The archiving institution is only providing on-site access to the ArcSys. However Jane Jones needs several days to do her analysis on municipal information. This means that she wants to save some of her "work files" to be used in the following days.

- She gets a personal ID and password for ArcSys purposes.
- She logs on to the ArcSys and uses it extensively.
- She bookmarks some of the pages she has found useful into a personal bookmark folder.

- She also saves some specific documents she has found useful for further purposes (e.g. images on a web site or image captures to be used in her study report).
- She also saves some of her performed searches for reference and possible re-searching.
- To access the materials she has saved she needs to be logged in to the ArcSys. Now being a regular user of the web archive she might also want to customize some aspects of the ArcSys User Interface.

Comments:

Type of user: Professional.

Technical requirements: Authentication mechanism. Personal book marking/saving facilities. A system for saving searches and performing them when requested. User-Configurable parameters for the UI. Guidance for these personified features.

Other: Priority medium - because all tasks could be done and it is convenient not a core functionality.

4.6 Genealogical Use Case

A genealogist frequently consults web sites maintained by other genealogical researchers. One day, a very useful web site containing a nominal index to an early national census returns a 404 – File Not Found Message. Over the next few weeks, repeated attempts to connect with the web site confirm that it no longer exists online.. The genealogist queries ArcSys which returns a list of dates on which the web site has been harvested; the valuable research materials will continue to be available to the genealogical research community.

Comments:

Type of user: Common citizen.

Technical requirements: -

4.7 Accessing ArcSys through another system (e.g. a federated search portal)

Jane Jones uses a federated search portal for finding different resources from a variety of databases. The web archive is defined as one target database in search portal. She searches the web archive as one of the targets in her search. After seeing the results (hit list) of her query she can select an entry and hop to the web archive user interface.

Comments:

Type of user: Any.

Technical requirements: An API and a common protocol between the portal and the web archive. Possibly authentication mechanism.

Other: Impractical in some scenarios. The use case raised a debate and the prioritization was postponed.

4.8 Search for copies

Two years ago, Jane Jones developed an open-source application and made it available on her own website. She encouraged other developers to download the code and make it available from their websites too. She is now interested in finding out to what extent people have actually done this. She uploads an MD5 fingerprint of the original file. ArcSys returns a list of domains from where an identical copy has been harvested.

For each domain, ArcSys displays the dates on which the copy have been harvested.

Comments:

Type of user: Professional.

The facility can also be offered by: Some search engines, to some extent. No one seems to be offering the service at present.

Technical requirements: Advanced, but not necessarily complicated/problematic.

4.9 Incremental perspective on the contents of a page

4.9.1 Main case

Jane Jones uses in her work a collection of articles on the address www.xtu.edu/poproj/papers/. Articles are added continuously but seldom. She keys in the URL of the list and her own e-mail address. Subsequently ArcSys sends her an email every time harvests find the page updated/changed.

Comments:

Type of user: Researcher.

The facility could also be offered by: Many others.

Technical requirements: Uncomplicated.

Other: It is important to note that because there is a long time span between each harvest of a page the users may have to wait several months before they are informed about an update.

4.9.2 Alternative case

Jane Jones is not interested in being contacted automatically, but when she browses the page she cannot see whether new articles have been added (the list is long, alphabetically sorted and new elements are not marked). Therefore she asks ArcSys if the page has been changed during the past months. ArcSys confirms this and returns a list of the few changes that have taken place since the latest harvest. ArcSys also informs about the date of this harvest.

Comments:

Type of user: Researcher/common citizen.

The facility can also be offered by: Unique for ArcSys.

Technical requirements: A sufficiently intelligent handling may be complicated.

Other: See above. This facility may be more useful than the one described above.

4.9.3 Alternative case

Jane Jones is a journalist and has found a press release on the Internet. The release is month’s old but it seems to match the present political situation conspicuously. She uses the verification described above in order to examine if the press release has been revised after it has been published.

Comments:

Type of user: Professional.

Other: As the above.

4.10 Providing persistent references

Jane Jones is writing a scientific article. One of the references is a web site with presentations of a number of relevant results. However, Jane Jones is not sure that this page will stay unchanged on the Internet while her article is being reviewed, edited and printed. The site may even disappear before a specific reader finds the reference. She keys in the URL of the main page and ArcSys returns a list of dates when the site has been archived. She consults the most recent one and finds it to be useful as reference. ArcSys provides an URL that refers directly to the persistent copy of the original site. She uses this URL as reference in her paper.

Comments:

Type of user: Researcher.

The facility is also offered by: Many others, but its value depends on the service provider's being a trusted repository [1].

Technical requirements: This scenario illustrates persistent identifiers. Dictionary creator asking for a short PI instead of the long (date+path) type.

Other: The need for this application is illustrated in [2].

4.11 On-demand harvesting

Jane Jones is writing a scientific article. One of the references is a web site with presentations of a number of relevant results. However, Jane Jones is not sure that this page will stay unchanged on the Internet while her article is being reviewed, edited and printed. The site may even disappear before a specific reader finds the reference. She keys in the URL of the main page along with her own e-mail address. As a response, ArcSys schedules a harvesting of the site. ArcSys informs Jane Jones by e-mail, when their site has been harvested, and again when the materials are available through its user interface.

Comments:

Type of user: Researcher.

The facility is also offered by: Many others, but its value depends on the service provider's being a trusted repository [1].

Technical requirements: This facility could be implemented by appropriate scripting around existing harvesters. The described model of interaction allows the archive to review the request before performing the actual harvest.

Other: The need for this application is illustrated in [2]. This use case needs more details in the description. More information will be added.

5 Cases relating to data mining

The ArcSys will in many cases become a huge repository and as such it will be a candidate for data mining applications. Basically the scenarios for these are infinite but one needs to draw the line, where "access" ends, and where actually the data mining and analysis starts. This can be difficult and the following cases focus on cases that are either immediately related to access techniques and specific index structures or are on the borderline between archive access functionality and analysis capabilities.

5.1 Data Analysis

5.1.1 Main case

Jane Jones is a researcher and has been assigned by the Danish Institute for International Studies to carry out a research on media and emergencies. She wants to analyze how the media responds to various emergencies - there is a tendency that the media responds more extensively to emergencies where Western people are involved e.g. the Tsunami in Thailand where many Western tourists were on vacation compared to other emergencies e.g. the earthquake in Bam, Iran. For this purpose, she has registered the list of the names of the websites of 30 international TV-channels in a personal profile. She now decides to study and analyze how the media has covered emergencies last year. The main task of each study is to state names of a selected number of specific disasters, which has taken place last year. ArcSys returns a graph illustrating how each TV channel associates with each disaster in the past year.

Comments:

Type of user: Professional.

The facility can also be offered by: Unique for ArcSys due to the time perspective.

Technical requirements: This facility could possibly be built on top of IA's Recall-interface.

5.1.2 Alternative case

In addition to the above analysis, Jane Jones wants to do a link analysis. Each of ACME's products has its own home page on ACME's web site, and she is interested in finding the pages that link to the specific products (especially if it is from pages related to interesting consumer groups). She keys in the home pages of the products and receives for each home page all incoming links that have been stored at the most recent harvest.

Comments:

Technical requirements: Technically uncomplicated, but the calculation could be time-consuming.

Other: As above.

5.2 Extract a subset of ArcSys for processing

5.2.1 Main case

Jane Jones is a researcher that wants to perform complex analysis, like computationally demanding analyses, on a larger subset of the archive by extracting a subset of the archive according to specific criteria. The criteria for subset selection may be simply a time-span or domain name, but it may also be more complex criteria such as language, file formats, time span, or other metadata either alone or in combinations. ArcSys returns extracted data that will be exported for processing elsewhere.

Comments:

Type of user: Researcher.

The facility can also be offered by: Unique for ArcSys.

Technical requirements: This requires a range of different indexes and a range of metadata attributes by which selection can be specified. The interface developed needs to return estimates of the size of the requested subset of the archive, potentially even distribution according to the selected criteria, as well as possibilities to select desired means of transfer (on-line after registration, DVD, HDDs), payment procedures and usage rights.

O*ther:* Because the data will be processed outside the web archiving institution great care must be taken to prevent unauthorized use.

5.2.2 Alternative case

Instead of exporting part of the ArcSys for processing elsewhere, the archiving institution may want to offer special APIs to allow direct, but controlled access to the ArcSys databases using the same methods as in the main case. This will safe every researcher to perform the same pre-processing steps. ArcSys returns extracted data that will be processed at the institution.

Comments:

Type of user: Researcher.

The facility can also be offered by: Unique for ArcSys.

Technical requirements: Same as for the main case, but instead of exporting the data a secure mechanism must be developed for allowing processing the data. In the extreme the institution may even offer a space from which code can be run locally on the archive, as some secure test bed sites allow.

5.3 Select a specific subset of ArcSys

Jane Jones is a researcher who wants to obtain articles on different types of technology, such as e.g. BLOGs. ArcSys returns a list of domains where the file has been harvested and for each domain the dates on which the file have been harvested.

Comments:

Type of user: Professional.

The facility can also be offered by: Unique for ArcSys.

Technical requirements: As blogs are not mentioned as such in the specific website texts, retrieval via certain document characteristics rather than keyword-based selection is required. This may require advanced selection interfaces such as selecting pages with a high number of incremental updates, certain layout or technological characteristics, such as the frequent occurrence of date fields, etc. Basically this implies a search over the structure of documents, rather than content/keywords.

5.4 Analyze the evolution of web technology

Jane Jones wants to analyze the evolution of Web technology like operating systems, web servers and versions, network connection speeds) in different domains. She supplies a list of domains that will be analysed. The ArdcSys returns a file containing the appropriate http response headers.

Comments:

Type of user: Researcher.

The facility can also be offered by: Unique for ArcSys.

Technical requirements: This requires access to log files that store http response headers, download times, transfer speeds and other relevant information needed.

6 Cases relating to presentation

6.1 Reconstruction of a lost web site

6.1.1 Main case

Jane Jones has established a web site, "jansen.org", with information about a number of peasants' lineage. However, the web site is lost when it is being transferred from one web-hotel to another. Jane Jones keys in "jansen.org" and specifies that she wants the latest version of the web site. ArcSys returns a ZIPped TAR-file with the most complete and updated version of Jane's web site as can be pieced together from the archive. When the file is extracted, directories and files are formed according to the structure the original web site reflected at the most recent harvest.

Comments:

Type of user: Common citizen.

Technical requirements: Uncomplicated.

Other: The users must be informed that ArcSys can only deliver static material – and not e.g. database tables or scripts.

6.1.2 Alternative case

Jane does not receive the contents of the web site but only a summary of the URLs in the archive. Jane can see the most recent date of harvest for each URL. She also receives a reference number. After having sent documentation stating that she in fact is the person Jane Jones who owned the domain at the most recent date of harvest, she receives the contents of the web site as stated above.

Comments:

Technical requirements: This alternative implies that the list of domains (including the identities of the domain owners) is stored.

Other: As above.

6.2 Blocking public access to Web documents

Jane Jones wants to ensure that certain documents providing personal information about her are being blocked from public access. She sends her request, accompanied by sufficient personal identification, to the administrating institution where the request is processed by local staff and a reply sent to Jane Jones. If the request is approved those pages which the user requests to be blocked from public access are marked accordingly.

Comments:

Type of user: Common citizen.

The facility can also be offered by: Unique for ArcSys.

Technical requirements: The system must provide an interface to mark those pages which the user requests to be blocked from public access. This implies performing a search with potentially elaborate heuristics returning documents mentioning Jane Jones name. In addition to keyword search for a name and abbreviations, including heuristics for middle name, the application potentially should differentiate between documents where the name is mentioned and such carrying potentially private information, such as CVs). A sufficiently intelligent handling may be complicated.

Other: In general, there might be a scenario where pages are requested marked for non-public display, or semi-classified information.

7 Use cases for archive internal use

Creating and maintaining a national or international Web Archive is a difficult and demanding task and a successful result will reflect the variety and complexity of the Web. The institution responsible for the Web Archiving activities (harvesting, archiving and access) will need data and information that will enable it to perform this task in the best possible manner. The following use cases illustrate some of the requirements for information and data that people in management positions will need. The use cases are divided in categories according to what function within the institution they apply to. Some are interdependent and some may serve more than one function.

The term "National Web" is used as a collective name for all the web sites that the institution has decided it should collect in order to fulfil its obligations according to its policy. This can e.g. be dictated by the Legal Deposit Law, institutional convention or practical considerations.

Many of the use cases refer to statistical information. It is not possible to rely on the data and information that can be collected and stored during harvesting, both because of the iterative and cumulative process involved, and because multiple harvesters may be used. Therefore statistical data and information must be extracted from the ArcSys and for that tools must be developed.

Many of the use cases depend on understanding the composition of the "National Web" and therefore it will be crucial for the Web Archiving institution to understand how the ArcSys reflects the real "National Web" or those parts of it that are harvested. In this context it is important to understand frequency of updates and the addition of new material, or completely new content.

7.1 Head of Acquisition/ Head of Legal Deposit/Collection Manager

7.1.1 Define ArcSys collection policy

The manager is planning the development of the ArcSys and the "National Web" collection policy. For this purpose the manager must analyse the "National Web", and define and identify segments that have certain common characteristics and features like: digital journals, topical websites (political, scholarly, etc), and media. For each of the segments defined the manager requests statistical data and information about: number of domains, size distribution of domains, number of web pages/documents, number of files, total size of the ArcSys and a list of Web-domains.

Comments:

Technical requirements: Statistical information, access tool like the NWA and data analysis. The data analysis can be complicated.

7.1.2 Schedule harvesting activities

The ArcSys Collection manager is planning the next year's development of the ArcSys in accordance with the institution collection policy (defined in 7.1.1). For this purpose the manager must schedule the Harvesting activities and in particular the frequency for harvesting different segments of the "National Web". The "National Web" segments have been identified/defined as being: the complete "National Web", digital journals, topical websites (political, scholarly, etc), and media (see 7.1.1). In order to set up a harvesting schedule the manager requests the following statistical data and information for each of the segments defined: number of domains, size distribution of domains, number of web-pages/documents, number of files, total size of the ArcSys, and a list of Web-sites

Comments:

Technical requirements: Statistical information, access tool like the NWA and data analysis. The data analysis can be complicated.

Other: This case may seem equivalent to 7.1.1and in some cases it may be. Use case 7.1.1 can be considered as a prerequisite for this case.

7.1.3 Inquiry about "National Web" composition or specific segments of it

A professor at a University faculty for political science asks the Web Archiving institution for a list of political web sites, i.e. web sites of official political parties and web sites that are primarily engaged in discussing political or general society issues. For this segment of the "National Web" (see 7.1.1) the manager requests a list of the Web sites, the total size of the segment and the number of web- pages/documents in the segment.

Comments:

Technical requirements: Statistical information, access tool like the NWA and data analysis. The data analysis can be complicated.

Other: Variants of this use case would be similar inquiries about cultural websites (specific, discussions etc) and about digital journals. Priority is medium because the archive will work without this functionality.

7.2 Head of Library (National Librarian) / Public Relations Manager

7.2.1 Composition of the ArcSys

A journalist asks the National Librarian to describe the "National Web". The National Librarian requests for each of the segments defined (see 7.1.1), statistical data and information about: number of domains, size distribution of domains, number of web pages/documents, number of files and total size.

Comments:

Technical requirements: Statistical information, access tool like the NWA and data analysis. The data analysis can be complicated.

Other: This is a variant of 7.1.3. Used on Swedish website.

7.2.2 Evolution of ArcSys year by year

The Public Relations Manager is compiling the annual report for the institution, and requests information about the most descriptive characteristics of the ArcSys, e.g. total size in Tb, number of domains collected, number of URL's, number of documents, and the most common formats. This information would apply both to the situation at year-end and to the yearly addition.

Comments:

Technical requirements: Statistical information.

7.2.3 Use of the ArcSys

The board of the institution wants to know how the ArcSys is used, i.e. how much, by whom and how. The manager requests information about the number of users, who they are (scholars, professionals, general public, etc.), what kind of domains they are accessing and the number of web-documents that are retrieved.

Comments:

Technical requirements: Statistical information provided by the Access system for the ArcSys and a user survey.

Other: Can be complicated and controversial to deliver. Not an immediate need, however, crucial for long-term success.

7.3 Manager for Information Technology / Data Base Administrator

7.3.1 Planning and Allocation of Computer Resources for harvesting and preservation

The managers responsible for information technology professionals and computer resources are making their budget plans and therefore must know what computer resources will be needed for harvesting, storing and preserving the ArcSys. The managers request information about the scheduled harvesting policy (see 7.1.2) and in order to project the evolution of the ArcSys: total size in Tb, number of domains collected, number of URL's, number of documents, and the most common formats. This applies to the current situation and the changes in previous years (see 7.2.2), and in addition the use of the ArcSys (7.2.3).

Comments:

Technical requirements: Statistical information.

Other: Scheduling policy as defined by the collection manager (see 7.1.2).

7.3.2 Control access to the ArcSys

The manager is defining how to provide access to the ArcSys in accordance with the legal environment for the ArcSys. Various user groups have been defined and the access rights for each have been defined. The manager needs information about where in the ArcSys metadata general access privileges are defined.

Comments:

Technical requirements:

Other: If the access rights depend on factors other than user category and access privilege as defined in the metadata, the ArcSys must be analysed with those factors in mind. This is complicated and unlikely to yield precise results. Awareness is given high priority and implementation medium priority.

7.4 Manager for Preservation of Digital Collections

7.4.1 Establishing a Preservation Policy

The manager is establishing a preservation policy for the ArcSys. The manager is aware that the distinctive feature of the ArcSys compared with other digital collections is the great variety of file formats. The manager requests information about what formats are in the ArcSys and the frequency distributions of the formats.

Comments:

Technical requirements: Statistical information.

8 References

[1] Research Library Group: “Trusted Digital Repositories: Attributes and Responsibilities”, Technical report, May 2002, http://www.rlg.org/longterm/repositories.pdf

[2] Rick Weiss: On the Web, Research Work Proves Ephemeral, http://www.washingtonpost.com/ac2/wp-dyn/A8730-2003Nov23?language=printer