APPENDIX 1

Archival Access Policy survey
Adrienne Sonder
June 19, 2003

The material below is excerpted and/or copied from the sources. Information was gathered from U.N.-affiliated organizations, and U.S. federal and state agencies. Findings suggest that few records at the state level have prescribed closed periods. The number is also small at the federal level. In international archives, the closed period varies depending on the organization and type of record. The shortest close period, imposed by the U. N., restricts records for a period of 20 years if the records were closed upon acquisition or if the confidentiality of the record is in question.

International organizations:

Selected guidelines for the management of records and archives : A RAMP reader. Prepared by Peter Walne [for the] General Information Programme and UNISIST. – Paris, Unesco, 1990.
Retrieved from:
http://www.unesco.org/webworld/ramp/html/r9006e/r9006e0q.htm#Access%20to%20the%20archives%20of%20united%20nations%20agencie s

[Ch. 25- Access to the archives of United Nations agencies]

6.2 **The State of Affairs**
Records/archives lacking standards and procedures for classification and declassification, retention periods, disposal policies and realistic conditions of access mean frustration to archivists as well as to internal and external users. The present survey has revealed a number of inadequacies in regard to the international organisations. Among the sample of the 34 international organisations chosen, only 41.2 per cent answered the questionnaire in a comprehensive manner. What is happening, if anything, in the field of archives administration in the other 58.8 per cent? So much information is missing that it seems almost impossible to get a clear picture of the actual situation….

..In general, rules and procedures relating to archives are rather scarce in the international organisations, although these instructions are essential in the maintenance of an operative records/archives management service. In this survey 11 organisations reported that they have such instructions, but only five submitted the texts as requested. Instructions of the IMF, UN Secretariat (also followed by UNECA and UNOG), UNESCO, UNICEF and WHO satisfy the standards of what is considered to be good archives administration. Otherwise, the so-called instructions are simply correspondence and registry manuals for secretaries, if the organisation has even such instructions….

6.3 **Diversity in Access**
Accepting the definition of access as "the availability of records/archives for consultation as a result both of legal authorisation and the existence of finding aids" means detailed responsibilities for archives administration. The manner in which UN agencies are dealing with this question differ in many respects and, for that reason, it is of interest to examine the content of selected rules and procedures.

6.3.1 **The United Nations Archives**
An Administrative Instruction, ST/AI/326, of 28 December 1984 "explains the guidelines concerning internal and public access to the United Nations archives". Access is given both to archives and non-current records kept by the service. It is clearly stated that staff members of the Secretariat may have access if they need the documents for official business, "except those subject to restrictions imposed by the Secretary-General". Regarding public access to archives and records, it is asserted that:

(a) they are open if they were accessible when created;
(b) they are open if they are more than 20 years and not subject to restrictions; and,
(c) they are open if they are less than 20 years and not subject to restrictions.

Consequently, the United Nations Secretariat follows a time limit of 20 years, but with flexibility in regard to non-restricted material. With respect to restricted records the Secretary-General has imposed two levels of classification:

- ST - Strictly Confidential to records originating with the Secretary-General, the unauthorised disclosure of which could "cause grave damage to confidence in the Secretary-General's Office(s) or to the United Nations".
- SG - Confidential to records originating with the Secretary-General, the unauthorised disclosure of which could "cause damage to the proper functioning of the United Nations Secretariat".

"SG - Confidential" records are automatically declassified when 20 years old, and "SG - Strictly Confidential" are reviewed for declassification at this age. Declassification in either case can be approved prior to the expiration of 20 years.

The United Nations Archives rule of a 20 year time limit is gaining wider acceptance, as in the case of UNESCO and UNICEF, and it could be a starting point in discussions on the subject of access.

6.3.2 UNESCO Archives

The "Rules governing access by outside persons to UNESCO's Archives" reveal that the holdings consist of documents, field mission reports and records. The first two are "freely accessible in the reading room of the Archives Section", although documents can be marked "restricted and confidential" and access given only "if the prior agreement of the relevant unit of the Secretariat has been obtained". Often, the documents are mimeographed or other multicopied material but not archival documents.

The third category, records, is another case. According to the Chief Archivist, a relaxation in access is currently under consideration, following the UN Secretariat's rule of 20 year time limit. Until any changes are made, the rules in force place it at 30 years, "with the exception of certain types of material where UNESCO may decide on a shorter period". A closed period limit of 50 years is imposed on the following material:

- files containing exceptionally sensitive information on relations between Unesco and its Member States, between Unesco and the United Nations, intergovernmental and non-governmental organisations;
- files containing papers which, if divulged, might injure the reputation, affect the privacy or endanger the safety of individuals;
- personnel files of officials or agents of Unesco; and,
- confidential files of the offices of the Unesco Director-General; Deputy Director-General and Assistant Directors-General.

It should be stressed that access to archives within the open period can be refused if they are "unmistakably of confidential nature still" and exceptions "to a paper or file that is not yet in the open period may be made by the Chief Archivist" after some provisions are fulfilled. The UNESCO rules thus also have a degree of flexibility.

6.3.3 UNICEF Records and Archives

This organisation has adopted rules and regulations similar to those promulgated by the UN Archives. The "Procedural guidelines for UNICEF records and archives" of 9 November 1983 follow closely the access conditions and 20-year rule adopted by the UN Secretariat. Archives and non-current records follow the same pattern of consultations and restrictions. Except that the latter can be imposed either by the Secretary-General of the UN, the Executive Director of UNICEF or their authorised representatives.

6.3.4 WHO Archives

These archives are defined primarily as "documents and correspondence of various kinds, received or produced by the Organisation .... in the course of carrying out its functions, and which have been

preserved in whatsoever form for documentary and historical purposes. External material, whether public or private, relating to the activities of the Organisation may be added to the archives; such material shall also be subject to these rules". That reference appears in "Rules governing access to WHO Archives" of 15 February 1974.

Access is given in situ after a time limit of 40 years but more recent material can also be freely consulted if it does not have any confidential component. In practice a pragmatic 10-year time limit is also employed. The determination of what is confidential is a prerogative of the organisation and is not clarified in the rules. WHO Archives also has material with closed periods of up to 60 years, i.e. "files containing information which, if disclosed, might prejudice the reputation, personal safety or privacy of individuals".

6.3.5 **IMF Archives**
This organisation applies no time limit for access to its holdings. "General Administrative Order No.26, Rev.I" of 1st November 1969, states: "All Fund documents and other records shall be considered restricted and not for public use except when designed for transmission to the public or specifically authorised for distribution to a particular recipient or group of recipients". The documents may also be classified as confidential or secret:

- "Confidential - records containing information, the unauthorised disclosure of which might be prejudicial to the interest of the Fund or its members. Records, the subject of which required limitations on use for reason of administrative privacy.
- Secret - records containing information, the unauthorised disclosure of which would endanger the effectiveness of a program or policy, or hamper negotiations in progress, or which could be used to private advantage. Use of this classification should be held to an absolute minimum".

6.3.6 **Overview**
In summary, from the above examples, it appears that access to the records/archives of international organisations is related to the identification of what is in the archives: the interpretation of the right to information:; respect for privacy of individuals; and the protection of the organisation's different spheres of interest. In addition, to open archives to the public means that the organisation must comply with basic requisites, including a good record management systemand the provision of user facilities. These goals have not been realised in many international organisations at the present time….

Guide to the archives of intergovernmental organizations. International Council on Archives. (Not dated). Retrieved online June 19, 2003 from:
http://www.unesco.org/archives/guide/uk/sommaire2.html

* *This site lists a number of international organizations with information on their archives administration policies. Listed below is a selection of those organizations that specified actual time periods to keep records sealed.*

1. **International Federation of Red Cross and Red Crescent Societies**

Access rules

The archival records are open to the public by appointment with the archivist, and in accordance with the following access conditions:
i. The Secretariat classifies as public the following records:
   a. Federation publications that the Secretariat makes available for sale to the public or distributes to the public for free;
   b. Decisions of the General Assembly, and policies or reports adopted through a Decision, except for those decisions, policies and reports designated confidential by the General Assembly.
   c. Decisions of the Governing Board, and reports adopted through a Decision, except for those decisions and reports designated confidential by the Governing Board.
   d. minutes and reports of statutory bodies more than 20 years old;

e. non-confidential files of the Secretariat that are more than 30 years old.

ii. The period after which a record becomes public is calculated from the date on which the record is closed.

iii. Records classified confidential, which are generally records containing personal data, are closed to the public.

2. **The Food and Agriculture Organization of the United Nations (FAO)**

Access Rules

The archival records of FAO are available for consultation by staff members in the course of their official duties in situ or on loan. Non-staff members, demonstrating a legitimate interest, may be given access to archival records, which have been closed for 15 or more years. In addition to the general 15 years closure period of records, special restrictions apply to records of confidential nature, such as personnel files of separated staff members, confidential reports, etc. The Director-General may, in appropriate circumstances and on recommendation of the Chief, Records and Archives Unit, remove these restrictions.

3. **International Committee of the Red Cross (ICRC)**

Access Rules

In January 1996, the ICRC Assembly adopted new "Rules governing access to the archives of the ICRC", which gave the public unrestricted access to archives dating from before 1950.
This historic decision was taken to respond to the desire of historians and many other people in search of accounts regarding individual victims of conflict and the conflicts themselves to extend the historical research undertaken since the late 1970s, at the initiative of the ICRC itself. A noteworthy example of this is a book entitled *Une mission impossible? Le CICR, les déportations et les camps de concentration nazis,* written by Professor Jean-Claude Favez of the University of Geneva, with initial publication in 1988 and a new edition appearing in 1996.
Extract from the "Rules governing access to the archives of the ICRC" of 17 January 1996.

"SECTION III: PUBLIC
Public archives Art. 6 : The general public has access to archives classified as public. The ICRC archivists select and make an inventory of archives to be classified as "public". After a set period of time, to ensure that such access will in no way be detrimental to the ICRC, to the victims that it is its duty to protect, or to any other private or public interests requiring protection.

Public archives Art. 7 :
[1)]Three types of document are to be found in the "public" archives :

- General ICRC files dating back more than 50 years, including minutes of the decision-making bodies;
- The minutes of the Recruitment Commission, the personal files of staff members and the record series containing personal or medical information dating back more than 100 years :
    - Access to biographical or autobiographical information on a specific individual is allowed after 50 years; such research, however, must be carried out by an ICRC archivist (see Article 10);
    - If permission is obtained from the individual concerned, the 50-year period may be shortened;
- Access to archival material from other sources which has been stored in the ICRC archives is authorized from the date set by the individuals or institutions that deposited the material at the ICRC.

[2)]The period during which the public is barred from consulting a file runs from the date on which the file is closed.

[3)]Documents that were open to consultation by the general public before being deposited in the ICRC archives remain so thereafter.

Special access Art. 8 :

[1)]The Executive Board may, before expiry of the time limits set in Article 7, grant special access to facilitate academic work which the ICRC itself wishes to see successfully completed or which it finds of interest.

[2)]The Executive Board adopts the *Rules governing special access to the ICRC's classified archives*. *Restrictions* Art. 9 : Public access to ICRC archives may be temporarily delayed in order to permit necessary conservation work to be carried out on the documents requested, or if no space is available in the reading room.

Fees Art. 10 : A charge is made for research carried out by ICRC staff at the request of persons outside the organization (see Article 7).

Use Art. 11 : No use may be made of the archives for commercial purposes unless a specific contract to that effect has been concluded with the ICRC."

With regard to access to the ICRC archives, see also Jean-François Pitteloud, "New access rules open the archives of the International Committee of the Red Cross to historical research and to the general public", in *International Review of the Red Cross*, September-October 1996, No. 314, p. 551-561.

State-level agencies

Parent and Child's Guide to Juvenile Records. Texas Youth Commission. (Last updated September 19, 2001). Retrieved online June 1, 2003 from: http://www.tyc.state.tx.us/programs/parentguide_records.html

> "In Texas there now exists a records system that is designed to limit access to your juvenile records after you reach 21 years of age if you do not commit criminal offenses after becoming 17 years of age. The system is called 'Automatic Restriction of Access to Records.' This is in addition to your opportunity to have your records sealed and destroyed under other provisions of the Texas Family Code."

The Voluntary Adoption Registry System. Texas Department of Health-Bureau of Vital Statistics. (Not dated). Retrieved online June 1, 2003 from: http://www.tdh.state.tx.us/bvs/car/open

> The Voluntary Adoption Registry system becomes open to adoptees, birth parents, and biological siblings once they are 18 years or older.

U. S. Federal agencies

Research Presidential Materials. National Archives and Records Administration (NARA). (Not dated). Retrieved online June 19, 2003 from: http://www.archives.gov/research_room/getting_started/research_presidential_materials.html#records

> "In 1978, Congress passed the Presidential Records Act (PRA), which changed the legal status of Presidential and Vice Presidential materials. Under the PRA, the official records of the President and his staff are owned by the United States, not by the President. The Archivist is required to take custody of these records when the President leaves office, and to maintain them in a Federal depository. These records are eligible for access under the Freedom of Information Act (FOIA) five years after the President leaves office. The President may restrict access to specific kinds of information for up to 12 years after he leaves office, but after that point the records are reviewed for FOIA exemptions only. This legislation took effect on January 20, 1981, and the records of the Reagan administration were the first to be administered under this law. Staff at the Reagan Library and Bush Library can provide additional information regarding access to Presidential records in their collections."

Rules of Access to the House and Senate Records. National Archives and Records Administration (NARA). (Not dated). Retrieved online June 192003 from: http://www.archives.gov/records_of_congress/information_for_researchers/rules_of_access.html

> "Although the House and Senate regularly transfer records to the National Archives and Records Administration, these remain closed to researchers for designated periods of time: 30 years for most House records, with investigative records and records involving personal privacy closed for 50 years; 20 years for most Senate records, with a similar 50-year closure period for sensitive Senate records. Some Senate committees have instructed the Center to open selected series of records to researchers upon receipt of the records by the National Archives and Records Administration.
>
> The records of Congress are not subject to the provisions of the Freedom of Information Act."

APPENDIX 2

Results of LANIC Electoral Observatory Exercise
(Internet Archive evaluation)

Survey of 148 URLs from LANIC's Electoral Observatory (http://lanic.utexas.edu/info/newsroom/elections/) run through the Internet Archive. The sample URLs covered a total of 21 elections held in Latin America between December 1998 and June 2002.

	Percent	Number
Not in Archive	13%	19
Dead Links	61%	91
Doesn't Cover Critical Period (of 129 sites)	36%	47
IN ARCHIVE:	**87%**	**129**
ACCESS PROBLEMS (of 129 sites in archive)		
No Access to Content	9%	12
Limited Access to Content*	16%	20
Less Limited Access to Content**	36%	46
Total Content Imperfectly Able to Access	**60%**	**78**
LINK LEVEL* **		
1st level links with some type of hindrance	7%	9
2nd level links with some type of hindrance	20%	26
3rd level links with some type of hindrance	13%	17
4th level links with some type of hindrance	9%	11
5th level links with some type of hindrance	1%	1
Total link hindrances**	**50%**	**64**

ACTIVE SITES (of 129 sites in the archive)	**%**	**Active**	**Total**
1999 elections	35%	22	62
2000 elections	34%	14	41
2001 elections	44%	4	9
2002 elections	100%	17	17
Total Active Sites	**44%**	**57**	**129**

* Both graphic and link problems prevent any but limited access to site content
** allows for access problems with link and graphics but with most of content captured
*** of 66 sites with limited or less limited access
**** of the 66 sites with imperfect access, there are 2 sites with only graphic problems but not link problems

APPENDIX 3

Nigerian Election 2003 Web Archive Links

Sites crawled from April 17- May 23, 2003

Election-Related Sites

European Union. Election Observation Mission to Nigeria 2003
http://www.eueomnigeria.org/
Independent National Electoral Commission of Nigeria
http://www.inecnigeria.org/
Nigeria First. 2003 Election
http://www.nigeriafirst.org/elections.shtml
United Nations Electoral Assistance Project in Nigeria
http://www.unnigeriaelections.org/

Political Party sites

All Progressive Grand Alliance. APGA Women
http://www.apgawomen.org/
All Progressive Grand Alliance Foundation
http://www.apgafoundation.org/
Alliance for Democracy
http://www.afrikontakt.com/alliance/
Alliance for Democracy (U.K.)
http://afenifere.virtualave.net/
Democratic Socialist Movement
http://www.socialistnigeria.org/
National Conscience Party
http://www.nigeriancp.net/
New Democrats
http://www.ndnigeria.com/
Nigeria. Presidency. National Orientation and Public Affairs (NOPA), Abuja - [Olusegun Obasanjo]
http://www.nopa.net
Nigerians for Good Governance
http://npgg.freecyberzone.com/
Peoples Mandate Party
http://www.peoplesmandateparty.org/

Presidential Candidate sites

Buhari, Muhammadu
http://www.muhammadubuhari.com/
Buhari 2003
http://www.buhari2003.org/
Buhari.org
http://www.buhari.org/
Buhari for President 2003
http://www.mbuhari.com/
Buhari - Okadigbo Campaign
http://buhariokadigbo.com/
Buhari-Okadigbo Campaign Organisation, UK & Europe
http://www.buhari-okadigbo.com/
Buhari / Okadigbo Campaign Organization - All Nigeria Peoples Party, ANPPUSA, Inc.
http://www.anppusa.org/

Nwachukwu, Ike Omar Sanda
http://www.ikenwachukwu.com/
Nwobodo, James Ifeanyichukwu
http://www.jimnwobodo.com/
[Nwodo] Chief John Nnia Nwodo, Jr.
http://www.johnnwodo2003.org/
Obasanjo, Olusegun
http://www.olusegun-obasanjo.com/
Okadigbo, Chuba
http://www.okadigbo4president.com/
Jibril, Sarah
http://www.sarahjibril4president.org/
Rimi, Dr. Mohammed Abubakar
http://www.rimionline.com/

Gubernatorial Candidate Sites

AKWA-IBOM

[Nkanga] Idongesit Okon Nkanga for Governor
http://www.hope2003.org/

ANAMBRA

Uzodike, Ajulu
http://www.ajuluforanambragovernor.com/

BENUE

Unongo, Wantaragh Paul Iyorpuu
http://www.unongo.com/

ENUGU

Nnamani, Dr. Chimaroke Ogbannaya
http://www.ebeano.org/
Aniagolu, Loretta
http://www.aniagolu.org

KWARA

Lawal, Mohammed
http://www.lafoga.org/

NASARAWA

Adamu, Governor Abdullahi
http://www.abdullahiadamu.com/
Daniel, Otunba Gbenga
http://www.otunbagbengadaniel.org/
Agagu, Dr. Olusegun
http://www.agagu.com/

Curatorial Monitored Nigerian Websites

The sites were first monitored by checking the front pages only and printing significant pages such as, front page, press releases, interviews and events. Starting May 7th, front pages and significant pages were checked for changes.

Political Parties

Site NAME	Site URL	4/15	4/16-17	4/23-24	4/29-30	5/7-8	5/13	Comments
APGA Women	apgawomen.org				checked	no change	no change	
All Progressives Grand Alliance (AGPA)	apgafoundation.org/				checked	no change	no change	
Alliance for Democracy	afrikontakt.com/alliance/				checked	no change	no change	
Allience for Democracy-UK	http://afenifere.virtualave.net/				checked	no change	no change	
Democratic Socialist Movement	socialistnigeria.org/				checked	changed*	no change	* front page
National Conscience Party	nigeriancp.net/		checked		no change	changed*	changed**	* front page ** no link to candidates
New Democrats	ndnigeria.com/				checked	no change	no change	
Nigeria. Presidency. National Orientation and Public Affairs (NOPA)	nopa.net	checked		changed	no change	no change	no change	
Nigerians for Good Governance	freecyberzone.com/				checked	no change	no change	
Peoples Mandate Party	peoplesmandateparty.org/				checked	no change	no change	

Out of 10 Political parties sites, 3 have changed. NOPA site changed after the elections.

Presidential Candidates

Site NAME	Site URL	4/15	4/16-17	4/23-24	4/29-30	5/7-8	5/13	Comments
Buhari, Muhammadu	muhammadubuhari.com/	checked		no change	changed*	changed*	no change	* front page
Buhari, Muhammadu	buhari2003.org	checked	changed	changed	changed	changed	changed*	* virus attached
Buhari, Muhammadu	mbuhari.com			checked	changed*	changed*	no change	* front page
Buhari, Muhammadu	buhariokadigbo.com/		checked	no change	no change	no change	no change*	*was down in the morning
Buhari, Muhammadu	buhari.org/pages/1/index.htm		checked		changed*	changed**	no change	* front page ** many changes
Jibril, Sarah	sarahjibril4president.org				checked	no change	no change	
Nwachukwu, Ike Omar Sanda	ikenwachukwu.com/				checked	no change	no change	
Nwobodo, James Ifeanyichukwu	jimnwobodo.com/		checked		no change	no change	no change	
[Nwodo] Chief John Nnia Nwodo, Jr.	johnnwodo2003.org				checked	no change	no change	
Obasanjo, Olusegun	olusegun-obasanjo.com/		checked	changed*	no change	no change	no change	*New presidential page up
Okadigbo, Chuba	okadigbo4president.com/intimation.htm	checked		no change	no change	no change	no change	
Rimi, Dr. Mohammed Abubakar	rimionline.com/				checked	no change	no change	

Out of 12 Presidential sites, 5 have changed. Every time Buhari2003.org was checked it had changed.

Gubernatorial Candidates

Site NAME	Site URL	4/15	4/16-17	4/23-24	4/29-30	5/7-8	5/13	Comments
Idongesite Nkanga	hope2003.org		checked		no change	no change	no change	
Ajulu Uzodike	ajuluforanambragovernor.com		checked		no change	no change	no change	
Wantaragh Paul Iyorpuu Unongo	unongo.com		checked		no change	no change	no change	
Osagie Obayuwana	nigeriancp.net/edo.html			checked	no change	no change	no change	
Femi Falana	nigeriancp.net/ekiti.html			checked	no change	no change	no change	
Chief Loretta Aniagolu	aniagolu.org		checked		site dead*	still dead*	site dead**	*files listed **no files listed
Dr. Chimaroke Ogbannaya Nnamani	ebeano.org		checked		changed*	no change	no change	*front page
Mohammed Lawal	lafoga.org			checked	site down	site up*	no change	*no change from 4/23
Adewunmi Abassi	nigeriancp.net/lagos.html			checked	no change	no change	no change	
Governor Abdullahi Adamu	abdullahiadamu.com			checked	changed*	no change	changed*	*many changes
Ogbeni Lanre Banjo	nigeriancp.net/ogun.html			checked	no change	no change	no change	
Otunba Gbenga Daniel	otunbagbengadaniel.org/			checked	no change	no change	changed*	*changes in the events section
Oyekan Arige	nigeriancp.net/ondo.html			checked	no change	no change	no change	
Dr. Olusegun Agagu	agagu.com/			checked	no change	no change	no change	
Oyebade Olowogboiga	nigeriancp.net/osun.html			checked	no change	no change	no change	
Femi Aborisade	nigeriancp.net/oyo.html			checked	no change	no change	no change	

Out of 16 Gubernatorial sites, 4 have changed.

APPENDIX 4

Timing Exercise

Measuring rates of change of Web sites based on typology

Using several tools, including the HTTrack Website Copier, the following 21 sites from Latin America were monitored for content changes during a two-week period (April 18 - May 2, 2003). The sample of 21 sites was designed to be as broad as possible in terms of both content (representing political views across a broad spectrum, ranging from mainstream groups to insurgencies, formal and informal groups, etc.) and form (mime types and file formats, small sites and large sites, etc.).

In terms of the typology used for this exercise, the following conclusions emerge:

- Party, candidate, and electoral coverage sites typically have regular or frequent content updates in the period leading up to the elections.
- Party sites for groups that are not engaged in a current electoral campaign are updated infrequently, with the exception of large, long-established, and well-endowed parties, like the PRI in Mexico.
- Sites that have a section containing news items, in this case including the alternative media and some of the New Social Movement sites, tend to be updated more frequently, in most cases daily or even multiple times during the day.
- New Social Movement sites tend to be more or less active in terms of content updates in relation to how current their "cause" is.

For each of the sites listed below, the number to the right of the site name is the number of days during the 10 day (M-F for two weeks) measuring period that the site contents were changed or updated.

Candidates & Electoral Coverage

All four sites in this category pertained to a "current" electoral campaign, in this case Argentina, with the election itself falling midway through the exercise period. As expected, the sites were very active: three of the four had content changes on a daily basis every single day during the exercise period. The Menem site changed on only two occasions.

- Kirchner
  http://www.kirchnerpresidente.com.ar/kirchner/ 10
- Menem
  http://www.carlosmenem.com/ 2
- La Nacion Suplemento Electoral
  http://www.lanacion.com.ar/coberturaespecial/lacarrerapresidencial/ 10
- UOL Suplemento Electoral
  http://www.uolsinectis.com.ar/especiales/elecciones/ 10

Political Parties

The four parties range across the political spectrum, and are a mix of "in power" and "opposition". Mexico's PRI party site had daily content changes. On the other extreme, Guatemala's FRG site had no content changes at all during the exercise period. The FMLN and FSLN sites are both very large and extensive, and had content changes intermittently throughout the measurement period.

- FSLN
  http://www.fsln-nicaragua.com/ 1
- FRG
  http://www.frg.org.gt/inicio.htm 0
- FMLN
  http://www.fmln.org.sv/ 4
- PRI
  http://www.pri.org.mx/principal/PRI.htm 10

Alternative Political Media
Both of these sites are very active; this exercise confirmed that the sites have multiple content changes and updates on a daily basis.

- Politica y Actualidad
  http://www.politicayactualidad.com/index.asp 10
- Argentina Centro de Medios Independientes
  http://argentina.indymedia.org/ 10

Insurgencies
The Mexican FZLN insurgency appears to be quite active; during the exercise period, content updates were registered on about half of the days. The Movimiento Bolivariano site is affiliated with the Colombian FARC guerrilla force; none of the pages on this site have been updated since November 2002.

- Movimiento Bolivariano Colombia
  http://www.movimientobolivariano.org/ 0
- FZLN Mexico
  http://www.fzln.org.mx/ 5

"New Social Movements"
Nine sites were chosen to represent this broad category. Two of the sites had no content changes at all during this period; five of the sites changed four times or less during this period; and two, which had a large amount of news coverage related to their area of interest, changed daily or nearly every day.

- Antiescualidos.com
  http://www.antiescualidos.com/indexnew.html 1
- Asamblea Popular Revolucionaria
  http://www.mbr200.com/ 1
- Chavistas.com
  http://www.chavistas.com/ 10
- Red Bolivariana
  http://www.redbolivariana.com/ 7
- NuevasBases.org
  http://www.nuevasbases.org/ 2
- Cordoba Nexo
  http://www.cordobanexo.com.ar/ 0
- El Corralito
  http://www.elcorralito.com/principal1.htm 0
- confinesociales.org
  http://www.confinesociales.org/ 4
- Asociacion Conciencia
  http://www.concienciadigital.com.ar/ 4

Detailed Results

Site	Last updated as of 4/18	21-Apr	22-Apr	23-Apr	24-Apr	25-Apr	28-Apr	29-Apr	30-Apr	1-May	2-May
http://www.kirchnerpresidente.com.ar/kirchner/	17-Apr	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
http://www.carlosmenem.com/	early April	no	yes	no	no	no	no	no	no	no	Yes
http://www.lanacion.com.ar/coberturaespecial/lacarrerapresidencial/	18-Apr	yes	yes	yes	yes	yes	yes	yes	yes	yes	Yes
http://www.uolsinectis.com.ar/especiales/elecciones/	18-Apr	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
http://www.fsln-nicaragua.com/	13-Apr	no	n/a	no	no	no	yes	no	no	no	no
http://www.frg.org.gt/inicio.htm	n/a	no	no	no	no	no	no	no	no	no	no
http://www.fmln.org.sv/	8-Apr	no	no	no	yes	yes	yes	no	no	no	yes
http://www.pri.org.mx/principal/PRI.htm	n/a	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
http://www.politicayactualidad.com/index.asp	17-Apr	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
http://argentina.indymedia.org/	18-Apr	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
http://www.movimientobolivariano.org/	prior to 2003	no	no	no	n/a	n/a	no	no	no	no	no
http://www.fzln.org.mx/	14-Apr	no	no	yes	yes	yes	yes	no	yes	no	no
http://www.antiescualidos.com/indexnew.html	17-Apr	yes	no	no	no	no	no	no	no	no	no
http://www.mbr200.com/	9-Apr	no	yes	no	no	no	no	no	no	no	n/a
http://www.chavistas.com/	18-Apr	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
http://www.redbolivariana.com/	15-Apr	yes	no	yes	yes	yes	yes	no	yes	no	yes
http://www.nuevasbases.org/	early April	yes	no	no	no	no	no	yes	no	no	no
http://www.cordobanexo.com.ar/	3-Apr	no	no	no	no	no	no	no	no	no	no
http://www.elcorralito.com/principal1.htm	28-Mar	no	no	no	no	no	no	no	no	no	no
http://www.confinesociales.org/	16-Apr	no	no	yes	n/a	yes	no	yes	no	yes	no
http://www.concienciadigital.com.ar/	15-Apr	no	no	no	no	no	no	yes	yes	yes	yes

6. Politics, law and economics

6.15 Politics and government
Back to hierarchical index

Government
Narrower Term
- NT1 Government policy (*National policy, Public policy*)
- NT1 Political institutions
  - NT2 Heads of state (*Presidency*)
- NT1 Public administration
  - NT2 Central government (*Federal government, National government*)
    - NT3 Civil service
      - NT4 Civil servants (*Public servants*)
    - NT3 Government departments (*Ministries*)
  - NT2 Governance
    - NT3 Electronic governance (*E-governance, Online governance*)
  - NT2 Local government (*Regional government*)
    - NT3 Municipal government (*City government*)

Internal politics (Domestic affairs, National politics)
Narrower Term
- NT1 Electoral systems
  - NT2 Elections (*Voting*)
  - NT2 Womens suffrage
- NT1 Parliament (*Legislature*)
  - NT2 Government control
  - NT2 Ombudsman
- NT1 Political crises
  - NT2 Political conflicts
- NT1 Political leadership
  - NT2 Politicians
    - NT3 Women in politics
- NT1 Political parties

Political doctrines (Political ideologies)
Narrower Term
- NT1 Anarchism (*Nihilism*)
- NT1 Capitalism
- NT1 Collectivism
  - NT2 Communism
  - NT2 Socialism
- NT1 Colonialism
  - NT2 Neocolonialism
- NT1 Conservatism (*Traditionalism*)
- NT1 Federalism
- NT1 Feudalism
- NT1 Imperialism
  - NT2 Colonialism
    - NT3 Neocolonialism
- NT1 Internationalism
- NT1 Liberalism (*Radicalism*)
- NT1 Marxism
- NT1 Militarism
- NT1 Nationalism
- NT1 Pacifism (*Antimilitarism*)

NT2 Conscientious objection
NT1 Pluralism
NT1 Regionalism
NT1 Separatism
NT1 Technocracy (*Meritocracy*)
NT1 Totalitarianism (*Authoritarianism*)
NT2 Fascism
NT2 Nazism
NT1 Utopia

Political movements
Narrower Term
NT1 Civil war
NT1 Guerilla activities (*Guerilla*)
NT1 Liberation movements
NT2 Womens liberation movement (*Feminism, Feminist movements*)
NT1 Nonviolence
NT1 Oppression (*Abuse of power*)
NT2 Resistance to oppression
NT1 Protest movements
NT1 Revolutionary movements
NT1 Revolutions
NT1 Riots

Political science
Narrower Term
NT1 Political philosophy (*Political ethics*)
NT1 Political power (*Executive power, Judicial power, Legislative power*)
NT1 Political theory
NT1 Politics (*Political development, Political life, Political reform*)

Political sociology
Narrower Term
NT1 Conflict research
NT1 Polemology (*War studies*)
NT1 Political behaviour (*Political attitudes, Political psychology*)
NT2 Political corruption
NT2 Political participation (*Public participation*)
NT1 Political communication

Political systems (Political regimes, Political structures)
Narrower Term
NT1 Colonial countries (*Colonies*)
NT2 Colonization
NT2 Decolonization
NT1 Democracy
NT2 Parliamentary systems
NT1 Dictatorship
NT1 Federation (*Confederation, Federal systems*)
NT1 Monarchy
NT1 Newly independent states
NT1 Republic (*Presidential systems*)
NT1 Self government (*Autonomous states*)
NT1 State (*National state, Sovereign state*)
NT1 World government (*World state*)

APPENDIX 6

Political Communications Web Archive:
Test Data Input Module for MODS Descriptive Metadata

Start Time

1 Title

2. Alternative Title

3. Name

4. Abstract

5. Capture Date Range

6.1 Subjects: Controlled Vocabulary - Geographical

a) Region

b) Sub-region

c) Country

6.2 Subjects: Controlled Vocabulary - Subject

a) Topic

b) Actor

c) Event

6.3 Subjects: Keywords From Site

7. Language

8. Genre

9. Access Condition

10. Active URL

http://

11. Archive URL

http://

12. Archive

Political Communications Web Archive

End Time

Submit Clear Form

APPENDIX 7

A South-North Perspective on Web Archiving

Discussion paper for the Meeting of the Curatorship Investigation Team of the Political Communications Web Archiving Project,
Austin, Texas, December 16, 2002.

Peter Lor
Hannes Britz

2002-11-08
Revised 2002-12-12

Introduction

The topic of this discussion paper is intellectual property issues relating to web archiving as seen from the perspective of indigenous groups. We intend first to clarify the relevance of these two terms to the Web Political Communications Archiving Project, then explore some legal and moral issues relating to the harvesting of web sites. We use the term "harvesting" as shorthand for identifying, collecting, organising, preserving and providing public access to web sites or parts thereof.

Indigenous groups

In the context of indigenous knowledge systems, "indigenous groups" usually refers to first nations, the mainly pre-literate original inhabitants of countries subsequently occupied by colonial powers, colonists, or settlers. Examples: the Inuit and Amerindians of North America, the Aborigines of Australia, the San, Khoi and Bantu speakers of Southern Africa. We do not think that this use of "indigenous peoples" is appropriate to the present project. "Indigenous" must be taken more broadly to cover peoples at all levels of technological sophistication. The people responsible for the web sites of interest to this project may be from ethnic, linguistic or other minorities, non-dominant political groupings or movements that are to a greater or lesser degree overshadowed or repressed by dominant groups. In our context we should use the term "indigenous" simply to mean "of the country" or "based in the country" or "originating in the country" concerned.

Furthermore we should look at intellectual property issues from the perspective of (a) the groups responsible for creating the web sites, and (b) the citizens or inhabitants of the countries in which web sites are set up. (Perhaps also for which they were set up. Web sites are not always operated from servers physically located in the countries concerned.) More specifically, we need to place the web sites in the context of national heritage. This implies that we should also take into account the interests of the national heritage institutions in those countries whose task it is to preserve the national documentary heritage and make it accessible in the long term.

Intellectual property

There are many forms of intellectual property and various rights thereto, not all of which are readily protected by western copyright law. In developed countries there is little doubt that websites are subject to copyright and are protected by copyright law (Harris 1998). In some developing countries the relevant legislation may not be clear on this, but in so far as they have acceded to the international copyright conventions the intellectual property of these countries will in the developed countries receive the same level of protection as that of those countries themselves. <<Look at international conventions?>>

As a point of departure I propose to assume that, with a few exceptions, all web sites we want to harvest are subject to copyright. I can think of at least two categories of exceptions: (a) cases where the creation of the web site may be considered an illegal activity by the government of the country concerned, and (b) cases where web sites are collected in terms of legal deposit.

It is a legal principle (at least in some countries) that illegal activities do not receive the protection of the law. Under repressive regimes certain political groups or activities may be outlawed. Hence the products of these activities -- their publications including their web sites, may not receive copyright protection. <<To be looked at more closely.>> This is a legalistic loophole of which institutions under the rule of law in a democratic country will not want to take advantage.

Legal deposit may have a bearing on the harvesting of web sites. I expand briefly on legal deposit because in the United States legal deposit is associated with, and may be confused with, copyright registration. Today in most countries and under international conventions legal deposit is not a prerequisite for copyright. Legal deposit is the obligation imposed by law on publishers, printers and/or other parties to deposit one or more copies of their products in designated places of legal deposit, in most cases the national library. In many countries, legal deposit legislation has been extended to cover non-print material. In some countries legal deposit now covers both discrete and online digital media, including electronic journals and also web sites. Depending on how their legislation is framed, in these countries it may be legal for a legal depository to harvest web sites without contravening copyright. This would not apply to other institutions within the country. Since legal deposit law cannot be applied extraterritorially, it would also not apply to institutions outside the country.

In South Africa, legal deposit extends to online electronic publications, including web sites. This is also true of Namibia. Unfortunately, the national libraries of South Africa and Namibia do not currently have the resources to implement the legislation. Legally, however we may require deposit and to achieve deposit we may harvest web sites – or so we think; some aspects of this still need to be clarified. However, the use of the deposited materials remains subject to our copyright law.

Leaving out the above exceptions, we can assume that all web sites are protected by copyright. Websites, as information in a tangible format, fulfil the criteria set for information to be protected by copyright legislation. These criteria are that the information must be in a tangible medium and must be controllable (Britz 1997:124). Thus harvesting them without the permission of their owners or creators is technically illegal. Why "technically"? Web sites, it is generally assumed, are put out on the web to attract as many visitors as possible, so to harvest a web site, i.e. download it to a server in order to preserve it, does not appear to be "wrong". It is, after all, for a good purpose. This brings us to the moral dimension of web site archiving.

Moral arguments

For the purposes of this section it is assumed that it would be very difficult for the creators of the political web sites of interest to the Project to monitor the Project's harvesting activities and take legal action against those carrying these out. The likelihood of being apprehended and punished is negligible. So let us assume that the fear of retribution is not a factor in our decision making. This disposes of the amoral argument that we can go ahead because we will not get caught.

Based on the natural law position it is argued that there is a strong relationship between morality and legality. This implies that moral reasoning can be used to critically evaluate, and in some cases reform, intellectual property regimes.

On the basis of this moral position it can be assumed that there may be circumstances in which it is morally justified to do something illegal. Two conditions apply. The first one is when a law is in itself immoral – for example discriminatory and oppressive legislation (as under the nazi and apartheid regimes). This might imply a moral imperative to disobey the law. The second is where there is moral justification to disobey a law. The moral justification is normally based on the outcomes of an action. For example, it may be illegal to stop and get out of your car on a freeway, but if you do so in order to assist someone who has had an accident, it would be justifiable. (Not all jurisdictions are so reasonable, of course. In the Netherlands a householder who had apprehended a burglar and locked him up in a closet to await the arrival of the police, was arrested and charged with unlawfully depriving the burglar of his liberty.) India (in 1948) as well as Pakistan (in 1975) ignored international copyright legislation to enable affordable distribution of knowledge (textbooks) to educate their citizens (Basung, 1984).

There is also a school of thought that strongly supports the view that information is a common good and that it is morally justifiable to distribute it for free. The motto "information wants to be free" often reflects the sentiments of this school (Himanen 2001). Strong support for this moral position on the free access of information comes from Barlow (…..) in his thought-provoking article "The economics of ideas" where he argues that the digitisation of information will bring about the end of intellectual property. The implications are clear: those who digitise their intellectual property will not be able to protect it – it will be 'a sinking ship' http://ifla.org/documents/infopol/copyright/jpbarlow.html

What then are the moral arguments in favour of archiving web sites without asking for permission?

(1) “It is a good thing to harvest web sites and make them available to political scientists, historians and others for study and research. Web sites are here today and gone tomorrow. If they are not harvested, they will be lost for ever. We are doing this in the interest of science.” Impressive. But in the interest of science too, the corpses of deceased aboriginal and native persons have been removed from their burial places, deposited in museums, and put on show in glass cases. There are limits to what may be done in the interest of science.

(2) “It is a good thing to harvest web sites and preserve them for posterity. They form part of a nation’s documentary/cultural heritage. More and more of our history is recorded in media other than print. Web sites are a particularly significant non-print medium. We owe it to the citizens of the countries concerned to preserve at least a representative sample.”

(3) “Web sites are part of the common heritage of humankind. Even if the citizens or institutions of the country concerned take no interest in their preservation, this should be done on their behalf in any case.”

(4) “In developing countries the institutions that should do this lack the capacity to harvest and preserve web sites. If we don’t do it, they will be lost for ever.”

(5) “In some countries the institutions that should harvest and preserve web sites are controlled by oppressive regimes. They may be prevented from carrying out this task or may be pressurised into introducing some sort of bias (e.g. bias in respect of selection and preservation decisions.). We can ensure that a representative sample of political opinion in the country is preserved and made accessible.”

(6) “Because web sites are so ephemeral, there is no time to approach the copyright owners for permission to harvest their sites. Communications with groups of this nature may be slow or erratic. They may be so preoccupied by their political struggle that they are not able to respond in time to request for permission. Their communications may be obstructed or monitored by an oppressive regime. There may be language barriers. They may refuse permission because of misunderstandings or for fear of manipulation, espionage or sabotage. We would be doing them a favour if we harvested their sites in spite of their objections. We are helping the powerless to get their message across.” The question arises, why should we assist this particular group to “get their message across”? Who decides?

There is an analogy to this altruistic impulse that steps over legal boundaries in order to do good: a non-governmental organisation such as Médecins sans Frontières might enter a war-torn area without the permission of the government of that country, to assist refugees in rebel-held territory. This seems like taking a noble risk. But there is another analogy. In previous centuries cultural treasures were taken from colonies and other less powerful countries and deposited in museums and other institutions in the imperial powers. The Elgin Marbles are a well-known case. Ex post facto such looting has been justified on the basis that, if they were left where they were, the cultural treasures would have suffered grave damage or been destroyed due to the negligence of their owners. Today this argument is widely rejected, and in many cases there are demands for the repatriation of cultural treasures. Clearly in this case altruism is a less credible motive.

A moral ‘good’ approach

The question can then indeed be asked: what is the morally correct approach to web archiving? The ‘free access’ argument in favour of web archiving can be as unjust as the solely economic approach favouring the strict control and use of information. The ideal position would be to find a moral balance between the ownership of information (property provisio) and access to information (access provisio). Such a position would correctly reflect the dual purpose of intellectual property design.

On the one hand it must be borne in mind that one of the basic moral imperatives of merit based justice (the Lockian view) is that creators and owners of information products and services (including those who create and own web pages) have a right to control their work as well as to be compensated (economically or otherwise) for it. On the other hand, justice, based on needs propagates the accessibility and fair use and distribution of those information products. This dual nature of intellectual property must be maintained – also with regards to web archiving.

A moral conflict can arise in those cases where the property provisio restricts the access provisio – as has been demonstrated in this paper. This implies that a choice has to be made between either access to

information or the ownership thereof, which might implies exclusion. In those cases where access to information services a societal goal, in other words where information products are seen as a common good, the moral argument must be in favour of the access provisio.

The question arises then: does web archiving serve a societal goal? As part of our national heritage it can be seen as a common good. Heritage is not only concerned with the past, but also with the future. In a sense, future use and enjoyment provide the only justification for the preservation of heritage. As we move into the future, the circumstances in which heritage materials such as political web sites saw the light recede into the past. No longer the subject of such intensely partisan goals and activities, these materials become of more general and scholarly interest, from where they can be integrated into a more balanced, mature and nuanced understanding of the making of a nation and its place in the world. It is in that sense that archived web sites will become national and ultimately international heritage. We further suggest that such an understanding is what makes heritage a common good.

Guidelines

Pragmatically, we are going to be in the business of harvesting at least some web sites without obtaining prior permission from copyright owners. If we were not, there would be little point in our being here. So, what to do if we "have to" harvest web sites without permission. The following guidelines are proposed in case where the property provision has to be overridden:

(1) Always ask for permission first.

(2) If this is not possible, ask for permission ex post facto.

(3) If permission is not granted, reconsider the continued retention of and access to the material. We should develop a set of criteria to guide us in these decisions.

(4) Make the material accessible to the original creators (i.e. the political groups or movements that set up the web sites).

(5) Make the material accessible to scholars and institutions in the country of origin. For them, lower barriers to access that are constituted by user charges.

(6) Take care not to play into the hands of repressive regimes.

(7) Take care that the interpretation of the material is not of such a nature as to reinforce First World prejudices about the countries of origin.

(8) Ensure that the interpretation of the material and research on it is not done only by US scholars, but also by scholars from the countries of origin. Encourage them to undertake research on the material by making available bursaries for postgraduate studies, stipends and visiting scholarships.

(9) Assist institutions in the countries of origin to build capacity (technological as well as methodological and ethical) to harvest and preserve their web material themselves in future.

<<Note: It would be interesting to look at these guidelines in the context of other, more general guidelines for ethical conduct in research, such as those of the African Studies Association (2002), which in short say:

- Do no harm
- Open an full disclosure of objectives, sources of funding, methods, and anticipated outcomes
- Informed consent and confidentiality
- Reciprocity and equity
- Deposition of data and publications

The guidelines proposed above for web archiving appear to be in line with ASA guidelines 1, 4 and 5, and partly with 2 and 3.>>

References

(Still under construction.)

African Studies Association. (2002) Guidelines of the ASA for ethical conduct in research and projects in Africa. Web page, URL: www.africanstudies.org/asa_guidelines.htm. Accessed 3 December 2002.

Barlow, J.P. 1994. The economics of ideas: a framework for rethinking patents and copyrights in the information age. WIRED (2.03, March 1994). Available at URL: http://www.ifla.org/infopol/copyright/jpbarlow.htm Accessed December 12, 2002.

Basung, L. T. 1984. Reprinting of foreign publications in some developing countries. Journal of Philippine librarianship. (March-September, 1984): 93-99.

Britz, J.J. 1997.

Harris, L.E. 1998. Digital property: currency of the 21st century. Toronto: McGraw-Hill Ryerson.

Himanen, P. 2001. The hacker ethic and the spirit of the information age. London: Secker & Warburg.

APPENDIX 8

Technical Challenges of Web Archiving

Leslie Myrick
November 14, 2003

As a complement to decisions concerning the selection, accession and management concerns that are the focus of Curatorial Team's Collection Policy, the primary Technical Collection Management decisions also involve capture, storage, preservation, management, and access. The requirements that inform all of these decisions are bound to be complex in the case of archiving and preserving born-digital objects; more complex still in the case of Web sites. Decisions must be made concerning harvest configuration and timing, data storage models and archive file formats; data format issues; preservation strategies – whether migration, emulation, refreshing, or some combination; data access mechanisms e.g. persistent identifiers; metadata standards and cataloguing; administrative access; user access mechanisms; and quality control. In a project such as the Political Communications Web Archive (PCWA), the Long Term Resources Management, Curatorial and Technical Teams' decision-making and construction of an architecture that will assure collection, preservation and access are intricately intertwined; we will lay out our evaluation of those technical aspects of the endeavor that can be separated out for scrutiny.

Because of the fluidity and complexity of the World Wide Web itself coupled with the volatility of the technology used to capture, store and preserve Web sites culled from it, a robust yet flexible architecture married to a metadata system that accounts for structural, descriptive, technical and administrative information is the key to managing these complex digital objects in order to assure their authenticity, completeness, long-term preservation and access.

Most harvesting projects/repositories embarking upon this task are invoking the OAIS reference model[1] and the Trusted Digital Repository model[2] as the twin pillars upon which to construct a viable system for preserving access to digital objects. An OAIS-compliant, trusted repository should be modular, scalable, and tightly bound by a flexible, extensible metadata system.[3] In such a repository five subsystems or entities assure preservation of ingested digital objects for the long term, and facilitate the smooth transition from SIP to AIP to DIP, with the end of rematerializing for users the archived digital objects stored in AIPs.

Many questions arise in the act of preserving digital material culled from the Web: what exactly is being archived? what are the "significant properties" that must be preserved? To what extent is it necessary to preserve the original look and feel in future access to the material?

The rhizomic nature of the Web makes the definition of the boundary of a Web site problematic – the great majority of sites point outward to other sites through hyperlinking. Should a repository turn off external links or leave them active? Archive external links or ignore them altogether? A repository's curators must distinguish those "significant properties" of a digital object that must be preserved; this could include or exclude external hyperlinks, "near files" such as stylesheets or icons that might not live on the same domain as the site, downloadable files in various formats, client-side scripting, dynamic functionality, etc.

A single Web page is a relatively discrete object with links to other HTML pages internal and/or external to the domain, along with inline or embedded video, sound files, images, graphics, ad services, stylesheets, javascripts, and perhaps database-generated material or other types of dynamic or deep Web content. And underlying the page itself is a substratum of code, whether HTML, XHTML or XML, or perhaps containing (and dependent upon) javascript, with accompanying stylesheets; or .cgi, .php, .jsp, .asp scripts or servlets, in the case of deep Web gateways. A Web archiving project should, wherever possible, archive all of the code and scripting that underlies the page. However, harvesting dynamic scripting may

---

[1] See the standard references e.g. the CCSDS Blue Book document *Reference Model for an Open Archival Information System*, 2002, http://wwwclassic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf;
Preservation Metadata and the OAIS Information Model, A Metadata Framework to Support the Preservation of Digital Objects, OCLC and RLG, 2002. http://www.rlg.org/longterm/pm_framework.pdf

[2] See *Trusted Digital Repositories: Attributes and Responsibilities*, RLG and OCLC, 2002. http://www.rlg.org/longterm/pm_framework.pdf

be a deep-Web pursuit that is, in most cases impossible. Most websters who use the robots.txt exclusion will undoubtedly protect the directories containing scripting from collection. This is one particular case where negotiation is necessary to preserve the site entire.

Another question Web archivists face at the outset is that of versioning. Because of the ephemeral nature of the Web and born-digital media, two sorts of versioning problems emerge in archiving: the capture and management of different versions of the Web page as modified by the creator or by a database, and the eventual refreshing of bits or the migration of archived pages by the archiving repository into different formats, either following a stated normalization criterion or as old formats become obsolete.

Undoubtedly many static pages remain static for their lifetime, but the percentage of pages modified daily or weekly, e.g. dynamic pages generated from databases or RSS feeds, is significant. Online newspapers are updated at least once daily, with the BBC perhaps at one extreme, claiming (if only because the homepage contains a dynamic datetime function) that the site is updated every minute of every day. According to a much-cited early study (2000) by Molina and Cho,[4] who crawled a set of 720,000 pages on a daily basis over four months, 40% of all Web pages changed weekly, and 23% of .com pages changed daily. Dennis Fetterly et al crawled 330 million URLs nine times over three months and found that half the sites changed weekly.[5] The greatest rates of change were found in pages that contained banner ads, counters and date scripts, news and stock ticker applets, or Weblogs. Jay Sethuraman et al found that 23% of their sample overall changed daily; while 40% of commercial pages changed daily. They set the half life for a Web page at ten days.[6]

Political Web sites, especially those belonging to radical groups and NGOs, are subject to spurts of activity around political events such as elections, coups-d'etat, legislation debates, and so on. Many of the URLs that will be monitored and archived by the CRL project are news portals and will thus undergo daily changes. In a similar vein, the online production of some radical NGOs might replicate the ephemeral nature of street pamphlets or graffiti. A particularly intriguing event that is surely be a target of an archiving project such as this would be the hijacking of a political Website. A typical eight-week-long snapshot crawl made by Alexa for the Internet Archive will happen upon such ephemeral occurrences only through serendipity. Is an Alexa crawl in and of itself sufficient for a selective, particularly volatile archiving project such as the Political Communications archive? Is it sufficient when supplemented by focused crawls as provided by the Internet Archive crawler, or HTTrack or wget run manually?

[4] J. Cho and H. Garcia-Molina. The evolution of the Web and implications for an incremental crawler. In *Proc. of the 26th International Conference on Very Large Databases*, Sep. 2000.
[5] Dennis Fetterly et al, A Large-Scale Study of the Evolution of Web Pages
http://research.microsoft.com/aboutmsr/labs/siliconvalley/pubs/p97-fetterly/p97-fetterly.html
[6] Jay Sethuraman et al., Optimal Crawling Strategies for Web Search Engines
http://www2002.org/presentations/sethuraman.pdf

APPENDIX 9

Digital Preservation Considerations for Web Archiving

Nancy McGovern
December 2003

The domain of digital preservation has matured to the stage where there are prevailing, if not universal, practices. At full maturity, the ideal would be to mange all digital objects that are selected for long-term preservation as identical objects, without regard to the file formats contained in the objects. For now, preservation approaches continue to be largely format-specific. At this point, there are accepted approaches for some file formats that have proven preservation track records; some good management techniques for other formats that are harder to preserve; and no known approach for some new, complex, or extremely software-dependent formats.

The file formats that are present on the majority of Web sites, for the most part, present fewer preservation problems than other types of digital collections because they are primarily text-based formats, mainstream image formats, or other widely-used formats. There are, however, application-dependent formats and other types of formats that do not yet have defined preservation pathways, and there will always be new formats for which preservation approaches must be identified. For this project, we reviewed prevailing practice, and considered the implications for Web archiving.

Appendix 34 provides detailed MIME results from a review of Web crawls for the test sets of Web sites used for the Political Communications Web Archiving project, specifically Asian and Nigerian, with comparisons to results for other test sets. All crawls showed the same top four mime types—text/html, image/jpeg, image/gif and application/pdf.—in the same order. Those four types represented 92.7%, 99.2%, 97.8% 97.6% of all mimes for the ARL, CURL, Asia and Nigeria crawls respectively. Nigerian sites showed an even smaller percentage of text/html objects, with over half the total mime objects being jpegs or gifs, by far the highest proportion of any of the crawls. (see the Mercator crawl results in Appendix 32)

Prevailing digital preservation practice

A digital archive has several possible options for accepting file formats:

1. Limit the file formats that will be accepted by the digital archive to a subset of formats for which the archive has established procedures that are affordable and/or doable.
   Considerations: This is a proscriptive approach that has the advantage of contingent the preservation activities of the digital archive to manageable options, but the disadvantage of decreasing the comprehensiveness of the digital archive collections.
2. Accept all file formats that are submitted, then seek preservation solutions for those formats that do not have a defined solution at the time of acquisition, treating all formats equally to the extent possible.
   Considerations: This is an inclusive approach that carries a potential risk for the digital archive organization that depositors or users will expect the digital archive to be preserving files for which the archive has yet to develop an implemented approach.
3. Accept all file formats submitted, then assign preservation level categories by formats to make explicit the extent to formats will continue to be available over time:
   e.g., this digital archive will provide full level 1 preservation for all text-based formats for an unlimited time period, and level 3 bit preservation for x type of application format for the next 5 years with review at that point.

   Considerations: This is an inclusive approach that, if done well, balances the capabilities of the archive and the expectations of depositors and users for accessing the files.
4. Accept all file formats submitted, then convert selected formats for which no preservation approach exists or that are not widely-used to one of a limited set of preservation formats as determined by the digital archive.
   Considerations: This is an inclusive approach that ensures the archive is able to preserve the files, but may entail loss of functionality that is not acceptable for the depositors or users of some collections. This may be especially problematic if the archive is not very explicit about what it will and will not preserve.

Whichever option works best for a particular digital archive, the organization that operates the digital archive should clearly and explicitly document the selected option(s) in its preservation policy, and make its policies and procedures for the digital archive widely available to depositors and users.

Implications for Web archiving

These preservation options have implications for Web archiving projects. Web crawls, the primary means by which a Web archiving project acquires materials, when successful, take in all of the formats that are present on a target Web site. After considering the options, the ideal for the Political Communications Web Archiving project would be to accept and preserve all formats as captured. The group determined that retaining the look and feel of the Web sites is a core objective of the project.

In practice that removes the first and fourth options described above. The ideal practice for political materials would not limit the intake of formats by type (option 1) or convert acquired files to more preservable formats (option 4).

Option 2 may be more problematic for a Web archive because the interaction with users (and depositors, when appropriate) will generally be through asynchronous global access using an interface to the archive. Therefore, it is more important for the preservation activities of the archive to be very clearly-defined and unambiguous to avoid unrealistic expectations.

Option 3 is a good match for Political Communications, and likely for other Web archiving projects. There are good examples of this approach that are already in place, e.g., the levels defined for the Sunsite at Berkeley (http://sunsite.berkeley.edu/Admin/collection.html), the Safekept approach of PADI (http://www.nla.gov.au/padi/safekeeping/safekeeping.html), and Harvard's Digital Repository Services (http://hul.harvard.edu/ois/systems/drs/policyguide.html#preservation). There is also ICPSR's Extent of Processing Approach (http://www.icpsr.umich.edu/help/abstract.html#EXTENT.PROCESS) that could be adapted.

In practice, this approach would:

- accept all formats that were captured by crawls
- categorize the level of preservation based upon the file format type, e.g., level 1: full preservation for text, HTML, XML, etc.; level 2: file migration and reduction of loss for GIF, JPG, PDF, etc.; and level 3: "as is" retention and monitoring for application files, proprietary formats, software-dependent files, etc.
- retain access to level 2 and 3 formats while actively seeking preservation solutions that would retain the look and feel of the original files
- document the level of preservation and update the preservation status over time for users of the digital archive

As the MIME results indicate, text and image files predominate, but there are still significant numbers of application, non-standard image, and atypical text files that might present preservation problems. The former is a boon for preservation, the latter the bane – and potentially very costly. We propose establishing a matrix of file formats using the MIME content types and subtypes to document the current preservation approach for each format, assign each format type to a preservation category, and define the level of preservation support by the digital archive associated with each category.

APPENDIX 10

Risk Management for Web Resources

Nancy McGovern
December 2003

A full risk management approach to the long-term preservation of Web resources requires a complex combination of organizational and technological quantitative and qualitative measures. Risk management protocols and techniques are well developed in many domains, wherever valued assets may be threatened by natural and man-made consequences, yet preservationists were fairly slow to engage in parallel developments.

Recently, risk management has become a trend in preservation, particularly for Web resources. There are two distinct varieties of risk concerns regarding the Web. The first defines risk based upon the potential liability of an institution based upon the content of its Web site, or a Web site for which it is responsible. The second defines risk based upon the potential threats to the integrity and longevity of a Web resource, including technological obsolescence, security weaknesses and breaches, human-error in developing and maintaining Web pages and sites, benign neglect, power and technology failures, inadequate backup and secondary systems. In this project, we are interested in the second classification of risk.

Similarly, Web archiving may refer to two distinct types of activities: one, monitoring and capture pertaining to Web-based publications, and two, capturing entire or portions of Web sites as discrete pages contained within a boundary defined by all or a segment of a URL (e.g., all of the pages at or beneath a specified directory level). This project is interested in both types of capture, though primarily the latter. The former can be viewed as a specialized part of the latter. It should be noted that risks to the individual publications may be easier to detect and prevent than to Web sites.

There are numerous ways to measure potential risks, but it is often a combination of factors that would identify real risks. The possible combinations that indicate risk to Web resources are not finite, but change as technology, institutions, and resources change. Perceived risk is also based upon an institution's determination of acceptable loss. Documented changes in the number or size of pages, structure, or format of Web pages and sites of interest may or may not indicate risk, depending on the context. Iterative crawls, ongoing monitoring, tools and techniques to detect and assess change, and increased familiarity with resources over time all for the development of risk categories and appropriate responsiveness. Risk can be measured in a number ways and at a number of levels, as discussed below.

Quantitative
Web crawlers and other tools can easily determine the actual size of Web pages and sites, and note incremental and sudden changes over time, but cannot determine the cause, nature, or impact of changes absent established rules or scales for interpreting the results. These kinds of quantitative measures may be the easiest to obtain, but are only reliable indicators in conjunction with other data or as analyzed within protocols and formulae that are appropriate to the level at which the change occurred. See Appendix 34 for detailed page number and site size comparative results for political Web sites.

Page-level
A page should be evaluated as a standalone entity for characteristics that might suggest good management or risk indictors, as well as within the context of a Web site. Was it created using current and prevailing markup languages, and metadata tags? Is it well-formed? Are there identifiable and meaningful dates associated with the page in HTTP headers and/or within the page headers, tags, or textual content? Are the MIME types for the page commonly used, open source, non-proprietary? Does the page contain any known potential weak spots for remote attacks based upon CERN reports and other security monitoring sources? Answers to these questions might provide indicators that pages are well-managed and likely to be maintained, or suggest that the pages are at risk.

Link-level
Link checkers and other tools will determine the status of internal and external links, and monitor changes in the status of links over time. Sudden missing links may be traceable to badly managed Web site redesigns or upgrades, either on the site through internal links or a linked site through external links, but

other kinds of changes and fundamental changes in the content that is linked to be the target site may be more difficult to detect.

Site-level
Web site crawlers, analyzers, and site mapping tools document the structure and physical content of a site (i.e., pages, attributes within pages), and track changes to the site over time. It is harder to detect the basis for those changes and to determine if the changes equate to risk because often these changes stem from organizational or technological change that acts upon the site. An understanding of events and other change drivers is essential for capturing sites of interest. This is particularly true for political Web sites, which have a higher incidence of event-based and topical Web sites. The Nigerian election Web sites provided a good example for study. See Appendices 32, 33, and 34 for results of crawls.

Server-level
Server-level changes can be harder to detect. For example, it is possible to identify the type of Web server software that a site uses, but, depending on the settings a Web master chooses, the specific version and other attributes of the software installation may be hidden from crawlers. Other significant pieces of information about the server configuration and its operation might also be hidden. From a system security perspective, these shields reflect reasonable measures to protect the site; from a risk management perspective that uses remote monitoring techniques, these protective measures may remove key indicators from monitoring, placing more emphasis on other characteristics that are more readily monitored.

Organization-level
Significant changes at the organization-level include reorganizations, major programmatic or mission changes, mergers, or dissolutions. These would occur at the administrative context and external environment layers of the Web site model presented in Figure 1. Archival organizations need to understand both to knowledgably capture Web sites. This is particularly true for political sites, many of which are event-prompted by elections, changes in government, and adverse actions by governments towards groups or individuals, etc.

Most of the technical characteristics listed above should be detectable on a local or remote Web site using available or tailored Web tools. There are organizational changes that would be significant risk triggers that are much harder or impossible to detect absent some way to retrieve and respond to events and updates. For example, a change in Web master, changes in page or site owners, and decisions about Web site management.

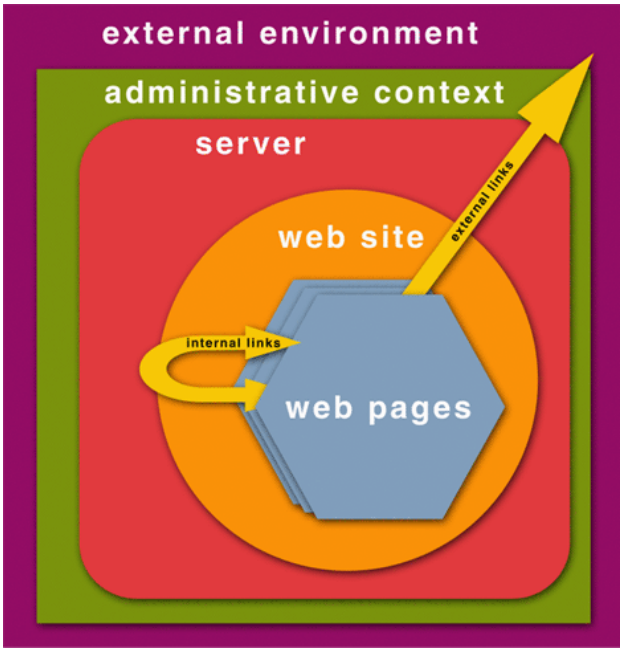

Figure 1. Virtual Remote Control Web Site Model[7]

Web site Typology

Web site typologies may be defined by content and purpose of the site, as described in the curatorial report, or by structural and other technical factors such as overall site size, number of pages, type of page, e.g. containing text-only, image-only, text plus image, dynamic indicators (forms, scripts, etc.), and incremental change over time by page and directory/location.

Using these factors in evaluating the Nigerian and Asian sites as examples of political Web sites, there are a number of characteristics that emerge (see Appendix 34 for more detailed results of the evaluation):[8]

- Generally, there are fewer pages per site than for other test sites. The other sites tend to be institutional, while the political sites are often more event or subject-based sites.
- The number of pages per site, particularly the Nigerian sites, tended to remain more stable over time. Similarly, the overall size of sites is smaller and, like the number of pages, tended to remain more stable over time. These characteristics might support less frequent capture cycles for categories of political Web sites, and might make it easier to define risk parameters based upon change in the number of pages.
- There is a higher use of Apache HTTP server software for Nigerian and Asian, particularly on the Nigerian sites, than is typical across the Web, a corresponding lower use of Microsoft, and a higher use of other software. The use of less well-known software might be worrying.

Frequency of capture

Establishing the frequency of capture for individual sites is a core element of a Web archiving program. These decisions determine the size and scope of the program, and contribute significantly to the cost of archiving. Some projects have opted to take ad hoc *snapshots* of sites. A risk management approach that features preliminary site characterization and ongoing monitoring and evaluation offers the potential to schedule selective and more appropriate capture. This necessitates a combined curatorial and technological effort. As implemented in a Web archiving program, frequency of capture would be influenced the following factors:

Objectives of the organization in capturing pages/sites:

- to fully document the site by capturing all changes to the pages/sites (identified by ubiquitous monitoring against previous capture/crawl for incremental change)
- to capture significant changes to pages/sites (identified by regular monitoring against previous capture/crawl - subjective piece: what is significant?)
- to record periodic versions of the site (set uniformly for all target sites or based upon some categorization using some assessment of change over time - knowing that some content will not be captured)
- to capture a copy of pages/sites as a oneoff (because the site is short-lived, the interest in it is low, etc.)

Level of interest by archiving organization: sliding scale from essential to collection(s) to of little value to collection(s). This may also be tied to the nature of control the archiving organization might have with the creator/producer of the target site, e.g., is there an agreement with the owner to archive the site or not.

Rate of change: there are quantitative (based upon numbers and sizes - generally easy to ascertain), and qualitative (based upon techniques to gauge changes in content - generally much harder if possible at all) measures of change. Monitoring sites for a control period often identifies individual pages, clusters of pages, or directories that have higher and lower rates of change.

Scale: size of pages/site, number of pages, complexity of site, type of formats of pages. Cost may be closely tied to this (storage and backup mostly, but other organizational costs)

[7] This is the model that Cornell devised in developing its risk management program.
[8] The test sites that Cornell uses include ARL sites, CURL sites, a sampling of commercial and government sites for comparison purposes, and the Asia and Nigerian Election sites. The observations here are not exhaustive, but they are suggestive.

Schedules for capture should be set and kept current based upon change indicators that are weighted by these factors, including objective, level of interest, and scale.

APPENDIX 11

Web Archiving Cost Issues (Technical)

December 2003

Overview

The Technical Team was tasked with looking into the costs of Web archiving. Supplementing our review, the Curatorial report discusses staffing costs for selection and data input. The Long-Term Resource Management report incorporates cost references into their discussion of archiving activity areas. We tried to look at overall costs with a particular focus on technical cost factors.

Our review confirmed that there is a lot of interest in, but not a lot of applied work on digital preservation cost models.[9] We opted to use a conceptual model developed by the Instruction, Research, and Information Services (IRIS) at Cornell University Library for the digital preservation management workshop (http://www.library.cornell.edu/iris/dpworkshop/) as a starting point, and to identify cost areas that needed more investigations within that model.

The Cornell model, which references Shelby Sanett's work, identifies three categories of cost:

> Startup Costs: usually one-time expenses, including technical infrastructure (hardware, software, networks), personnel and services, and institutional overhead.
> *Notes for PCWA*: Technical infrastructure costs for collaborative Web archiving are influenced by the choice of crawler (e.g., cost, human resources needed to operate, server capacity required to run, storage considerations of output); the spread of personnel across the collaborative (e.g., location, seniority); overhead factored based on participating members plus central unit, if appropriate.
>
> Ongoing Costs: costs for maintaining once established (equipment, services, staffing, overheads).
> *Notes for PCWA:* Shared costs of a collaborative may help lower these because not every member may have to sustain all of the categories.
>
> Varying Costs: unanticipated – resulting from a major technological change, disaster of some kind, incorporation of a new preservation approach, new formats to preserve, or unexpected additions to the digital archive
>
> *Notes for PCWA:* All of these could occur in a Web archiving collaborative.

Startup and ongoing costs are by definition more quantifiable. The Curatorial report contributes quantifiably to both categories; the Long-Term Resource Management report qualitatively. We identified two specific areas (with both startup and ongoing costs, plus the strong potential for varying costs) that required data gathering for our evaluation: storage and staffing. In addition, the methodology evaluation considers program costs and the harvester evaluation includes cost implications. We acknowledge that even open source crawlers have associated human, equipment and other costs to incorporate.

Storage Issues and Costs

A federated storage model with a primary centralized storage site supplemented by redundant mirroring of some content, and local archiving of branded content would require a range of storage systems. A high end digital library server to handle a project, such as Web archiving, that is both resource and storage-intensive would be best represented by a Sun 15K with at least of 10 Terabytes of disk storage (add Storedge as necessary + tape backup); and 24-104 CPU. A Sun Center of Excellence 15K package comes with a price tag of around 4 million dollars. The Kulturarw3 server configuration is one Sun 450 for harvesting and another for storage/archiving. They use an AML/J Tape Robot for mass storage and a 1.5 Terabyte disk array as a disk cache.

[9] For example, Brian Lavoie of the Research Department at OCLC has produced an interesting model based on roles for archiving: http://www.oclc.org/research/projects/digipres/incentives-dp.pdf. Shelby Sanett has framed a discussion around cost factors for preserving electronic records is adaptable to other contexts: http://www.rlg.org/preserv/diginews/diginews7-4.html#feature2.

Disk storage is imperative for preserving digital assets; tape should ideally only be used for emergency backup. Fortunately the trend has been for disk prices to decrease yearly by a factor of 2. Cheap disk at present can be easily negotiated at $3,000-$10,000 per Terabyte. Most large projects canvassed to date maintain around a 20 Terabyte target for available disk.

Technical Staffing Requirements

The technical team undertook an informal email survey of a decade of projects in production, testbed and planning stages to canvas their representatives on staffing requirements, both actual and projected (See Appendix 12). What is surprising is the patent understaffing of most projects, especially in the IT realm. The survey also makes some interesting revelations about the integration or lack thereof of librarian/ archivist/curatorial/project manager personnel and IT personnel. Many projects, especially in Europe and Australia, appear to be weighted heavily in one direction or the other. This weighting has interesting repercussions on collection policy, management, metadata valuation and modes of access.

Assuming that the group managing the digital archive is not working in a vacuum, but is a component of a larger digital library structure, a minimally comfortable staffing configuration for an ambitious, selective political communications web archive would probably look something like this:

- 2 programmers/engineers initially for application development, R & D, access mechanisms; 1 programmer after the initial set up and q/a period.
- .10 system administrator; .05 network administrator; .10 DBA, .05 systems backup and recovery
- 1 Director; 1 Project Manager; 2-3 of Librarians/Curators/Archivists

Factors that will affect staffing (and costs):

- Scope of the archive: legal deposit for a national domain vs. a selective archive.
- Nature of collection/harvest: broad crawling vs. selective and/or legal deposit; push vs. pull; surface vs. deep web.
- Whether harvest/deposit is negotiated with the creators/website owners
- Extent of curatorial input in preselection vs. automated seedbed crawling.
- To what extent the material is catalogued or otherwise made available for discovery.
  MARC? EAD? METS?
- Quality Control Methods
- Nature of the archival storage
  Storage, maintenance, preservation: physical and logical
  Refreshing/Migration/Emulation strategies
- Whether software is open-source/homegrown or uses proprietary software with service contracts.
- Access control
  Nature of access interfaces.
  Indexing, search and retrieval; special interfaces e.g. Wayback Machine with timeline functionality.

APPENDIX 12

Survey of Staffing Requirements/Technical Expertise in Related Projects

The following is a survey of ten web-archiving projects, roughly half using broad harvest and/or some combination of push and pull legal deposit collection, with the other half using selective harvesting. Among the selective projects four are involved in collecting and archiving political or governmental websites. Factors to consider in assessing personnel required, such as the extent of selection and cataloguing are mentioned where known.

Legal Deposit/National Libraries

PANDORA – Project Manager, 4 librarians; 1 IT expert; systems administrators; 2 FTE preservation staff are investigating long-term preservation issues.
They make a selective crawl and catalog for the most part at the website level in MARC.
http://pandora.nla.gov.au/index.html

Kulturarw3 – Staff is made up of 2 fulltime IT staff + 1 parttime IT staff
They undertake an automated broad crawl of the entirety of the Swedish web and do not catalog at this time. Full-text indexing is the intended means of discovery.

http://www.kb.se/kw3/ENG/Default.htm

WARP – employs 4 Librarians. They are in negotiations to outsource IT to a large vendor.
This is a deposit library in its early stages.
http://warp.ndl.go.jp/

Political Web Archives

MINERVA – Project Coordinator + 1 Librarian + 2 IT staff fulltime; part time: 2 cataloguers; 2 IT staff.
There is an expectation that more IT staff will be devoted to the project in the near future. They are in consultation with NDMSO on cataloguing and MODS issues. They contract much of the actual cataloguing out to WebArchivist.org; there is considerable consultation with the Office of General Council and Office of Strategic Initiatives.
http://www.loc.gov/minerva/

PRISM – Manager + 6 staff made up of 2 librarians, 2 researchers, and 2 programmers.
This is a selective archive that monitors risk factors for political websites in South East Asia and elsewhere.
http://www.prism.cornell.edu/

LANIC – 1 Director, 2 project and content managers, 1 programmer or IT specialist, and 5 half-time student research assistants.
This portal for Latin American political websites uses selective crawls and offers a category-based browse interface for access to web materials they curate.
http://lanic.utexas.edu/

Netarkivet.dk – 3 participants from the State and University Library (of Denmark), 2 participants from the digitization and web department of the Royal Library, Copenhagen; 4 participants from the Centre for Internet Research, University of Aarhus, comprising two professors and two MA student assistants.
http://www.netarkivet.dk/index-en.htm

Archipol – 1 Project Manager + 4 technical staff
http://www.archipol.nl/english/project/

Miscellaneous Archives

Wellcome Institute Medical Web Archiving Project – 1 Technical Developer; 1 Librarian for Selection, Cataloguing, deploying PANDAS.
http://library.wellcome.ac.uk/projects/archiving.shtml

Internet Archive – for broad crawls they outsource to Alexa, where a team of three handles the crawling: 1 Sr. Development Technician; 1 Crawl Engineer; 1 Test Engineer; for large-scale focused crawls using their own crawler a similar group would be employed. They also employ 1 Data Archivist to maintain and preserve the data and any interfaces; 1 Systems Administrator; and 1 tools developer.
http://www.archive.org/

APPENDIX 13

Comparative Merits of Current Methodologies

Leslie Myrick, November 14, 2003

Note: This summary does not include some results that were produced after the November 17-18 meeting at the Library of Congress, including the most recent information about the Internet Archive open source crawler and final results from the Nigerian crawls

Armed with the issues and questions presented in Appendix 8, the Technical Team began its investigation of current collection methodologies and preservation programs by examining two of the larger and more successful National Deposit Library Web-archiving programs: the National Library of Australia's PANDORA project and the Kulturarw3 project at the Royal Swedish Library. (A fuller version of this evaluation can be found in Appendix 14.) During our planning stage the star of the Web-archiving project of the Japanese Diet Library appeared to rising, so we pursued interviews with them as well and include a full evaluation in Appendix 15. Our Web-archiving advisor and content provider for the evaluation, the Internet Archive, especially insofar as it has partnered with the Library of Congress' MINERVA project, will be treated separately below after an initial comparison of PANDORA and Kulturarw3 in terms of storage, preservation, metadata and access issues.

Our preliminary evaluation of the national deposit libraries PANDORA project and Kulturarw3 served to sketch out two diametrically opposed approaches to collection, cataloguing, management and access; the subsequent evaluation of IA/MINERVA serves as a sort of dialectical completion insofar as the MINERVA project has consisted of a National Library (of sorts) doing both selective harvesting and culling broad swath content provided by the Internet Archive's own focused crawler and the Alexa crawler respectively.

Although the PANDORA project and Kulturarw3 are both actively involved in legal deposit collection, these Web-archiving projects' methods and practices stand at different ends of the Web-archving spectrum on many fronts. Their differing practices reflect their respective collection and collection management policies as they relate to curation, selection, management, preservation and dissemination of Web-based materials.

While PANDORA negotiates relationships with all of its content creators and uses a selective approach to capture, supplemented by a fair amount of push technology from the creator to the archive, Kulturarw3 undertakes primarily a series of broad swath automated snapshots of all that is deemed "the Swedish Web". For crawling, PANDORA has developed in-house a harvesting/cataloguing application aptly named PANDAS that is based on Java WebObjects wrapped around the popular and free offline browser HTTrack as its harvester. The Kulturarw3 team for its part has adapted the software for the open-source Combine Harvester, originally more of a Web indexer than archival harvester, to its collecting needs.

The latter is arguably a more robust collecting application along the lines of the NEDLIB harvester, using a series of daemons or java classes to automate a number of complex harvesting tasks. On the other hand, the cataloguing modules built into the PANDAS system satisfy the particular focus of the PANDORA archive, which is to collect primarily discrete Web documents that can be catalogued in MARC and entered into their OPAC. At this point in time Kulturarw3 does not create library catalog entries for their Web material, but will depend on full-text search against the archive. This dichotomy between the librarian-centric approach to discovery and the IT-centered approach will be addressed more fully later in this report when we consider metadata and access.

Storage and Management of Archived Materials

The NLA's PANDAS implementation of HTTrack creates a mirror of the Web site and a series of logs containing a subset of HTTP header information and crawler tracking information that can be mined for descriptive and preservation metadata. Captured mirrors are stored on an HSM file system with tape backup. An original archived copy is stored, and separate work and display copies are created. Kulturarw^3 uses a multipart MIME file to hold the collection metadata, the header metadata and the file/object itself, very much like a typical output of GNU wget with headers written into the top of the file. They also store several files together in aggregates, similar to .Alexa .arc file aggregates. For storage they deploy an HSM system with 20 TB-capacity storage on DLT 7000 tape complementing disk storage of 1.5 TB.

Data Format Issues

Although the NLA has observed with interest the National Archive of Australia's practice of prescribing a limited number of file formats that it will accept into their archive, PANDORA is not planning to restrict or normalize its MIME types at this time. Because something approaching 90% of the data belongs to one of the four or five most common MIME types they plan to use preservation strategies including migration and emulation to deal with the odd 10%. More than one format is, however, kept of documents originally harvested or deposited in XML or PDF or Word in order to future-proof their continued survival.

The Swedish Library had over 400 MIME types registered in 2001; it is presently collecting close to 800 MIME types. Roughly 90% of them belong to the five most common file types. They plan to use migration and refreshing, rather than emulation, to preserve them.

Long Term Preservation Strategies

Three essential levels of preservation have been canonized in studies such as the interim report for the MINERVA prototype[10] as: the preservation of bits; of content (objects) and of experience (look and feel), with a rising scale of cost and labor-investment. Similarly a triad of preservation strategies is considered by most digital archiving projects: the refreshing of bits to new media; migration to other formats or other media as they become obsolete; emulation of original soft- and hardware environments and actual soft- and hardware museums.

PANDORA is planning to wield the full range of preservation strategies: migration, emulation, hard- and software museums, or just plain refreshing for data that cannot be otherwise migrated or emulated. Kulturarw[3] is depending wholly on storage and migration; they do not plan to use emulation or hoard obsolete soft- or hardware.

Metadata

Overarching management issues are the use and promulgation of standards; and where possible the adoption of open vs. proprietary standards in the operating system, software, markup, and metadata. Metadata must be standardized to allow interoperability in the case of distributed archiving systems, and it is generally good practice insofar as metadata usually ends up serving as both a management/preservation tool and an access tool. XML has become the lingua franca of not only data transfer but data and metadata management; metadata schemata such as METS, MODS, MIX and the imminent PREMIS preservation metadata schema are taking advantage of the standards-based interoperabilityof XML encoding.

An emerging issue in the face of descriptive and technical digital object cataloguing costs that are projected to be prohibitive is the feasibility of the automated extraction of descriptive and preservation/technical metadata from the assets themselves along with server-delivered HTTP headers that record the client/server transaction, and any additional file headers created by the harvesting module (see Appendix 20). The harvesting application should include filtering modules that can extract a good subset of metadata from the captured material itself. In the case of the Alexa/IA metadata output, the technical team wrote a series of post-processing scripts to process the.dat or metadata file that accompanies each .arc and dump the metadata into a database. It also scraped additional metadata out of the archived files – Web pages and binary files -- themselves.

A related issue in automating the population of metadata databases with programmatically extracted metadata is to what extent creator-generated metadata such as <meta> tags and <title> tags can be trusted. Before we had cracked open our first .arc file we sent some simple perl LWP crawlers after the Web sites on our seed URL lists to ascertain how many sites used <meta> DESCRIPTION and KEYWORDS tags, and how they used them. We made a similar survey of <title> tags in HTML headers. The results can be seen in Appendices 29 and 30 respectively.

The misleading nature of creator-generated metadata can result from the desire to manipulate search engine ratings or from mere carelessness or error, as we show in yet another appendix, entitled the Case

[10] WEB PRESERVATION PROJECT: INTERIM REPORT by William Y. Arms, January 15, 2001 http://www.cs.cornell.edu/wya/LC-Web/interim.doc

of the Purloined Metadata. Here, a Web page creator for a French Marxist online journal, in copying a javascript from a German sports-related Webpage, also imported the <meta> tags belonging to that page: <META CONTENT="Sport sports Baseball Basketball Beach-Volleyball Bob Boxen Bundesliga Bundesligavereine Championsleague DEL DFB DFB-Pokal Eishockey Ergebnisse Europameisterschaft Europapokal Fernsehen Football Formel1 Formel3 Fußball Golf Hallenmasters Handball Hockey Inline-Skating Leichtathletik Motorbike Motorrad Motorsport Nationalmannschaft NBA NFL NHL Reiten Rodeln Schwimmen Skifahren Skispringen Snowboard Sportarten Sportnachrichten Surfen Tennis … [many other terms deleted here]" NAME="keywords">. (See Appendix 31 for the full account).

The NLA PANDORA Project depends on MARC catalog records for discovery but has made provisions for the collection of a sizeable subset of preservation/technical metadata for long-term preservation of their Web assets. In the PANDAS system, all metadata is processed along various points of the selection/collection/preservation continuum. Collection metadata is automatically harvested from HTTP header files, but each title is described in MARC by a human cataloguer. There is adequate control over digital provenance also built into the system; e.g. any changes to the Web site made through human intervention is manually added to digital provenance metadata.

Kulturarw^3 depends on the automatic generation of collection metadata from the crawler. This extracted metadata would include any information provided by the server-delivered HTTP headers, e.g. Last Modified Date, along with metadata about the capture event provided by the crawler itself. Having opted for full-text search against the pages themselves, they do not enter MARC cataloguing data into their OPAC for each title at this time.

The MINERVA/Internet Archive/WebArchivist.org Synergy

In some ways having constructed a hybrid of the harvesting methodologies of the two approaches outlined above, the MINERVA Project team, working for the most part in collaboration with the Internet Archive and WebArchivist.org, has been responsible for researching and developing tools to collect and archive born-digital objects from the Web into a series of Internet Libraries. This is an event-driven project with primarily a political focus, with discrete projects archiving Web sites that cover Election 2000, Election 2002, The 107th Congress, September 11th, the Iraqi War, and Winter Olympics 2002. For the reasons adumbrated above their project can be seen as a rich source of information about other facets of political Web archiving that were not entirely visible in the previous studies.

Harvester

The MINERVA prototype used the offline browser HTTrack run manually against twelve sites to create its testbed. For the subsequent event-based Internet Libraries they use a combination of raw Alexa crawls and focused Internet Archive crawls for their material. Alexa crawls were seen as problematic in terms of timing – many crawls had to be performed to collect an entire site. Alexa can take many days to capture a site (as we've seen from our data) and uses a breadth-first algorithm, which tends to make deep level crawling problematic. There is also no mechanism to alter crawls once they have been seeded to produce desired results.

Coverage, Scope and Size

The Election 2000 site is hosted on IA servers; 800 sites were archived daily between August 1, 2000 and January 21, 2001. It contains 800 GB of data, with 72,135,149 valid original objects. After de-duping 59,429,760 duplicate objects were removed (!); some 12.7 million valid objects remain, of which 9,972,695 are unique objects (according to checksums). User Access Mechanisms include a subject-oriented directory, where you can select sites by subjects such as Green Party Sites, Humor and Criticism Sites, along with the WayBack Machine method of choosing a URL and selecting one of many archived versions of the site.

For the September 11th Collection, they crawled daily for 3 months. The original archive contained 5 TB of data, that was honed down to 1 TB after the Internet Archive performed de-duping. It originally contained

331,299,192 objects, pared down to 55,224,374 unique objects after checksum analysis. The interface was developed by WebArchivist.org. Sites can be browsed according to four topic headings; or by a full index of sites.

Election 2002 consists of 4,000 sites archived between July 1, 2002 and November 30, 2002. The initial release consists of sites belonging to congressional and gubernatorial candidates; it will be expanded to include party and interest group sites.WebArchivist.org designed the search interface, which would be more correctly labeled a browse interface. Options now include not only browse by category but also alphabetically by state or candidate's name. The LOC would like more search capacity, and evidently WebArchivist.org is working on a searchable metadata database. An inhouse index of the archived home pages has been undertaken using Inktomi, but that does not seem to be the final solution.

Storage and File Formats

The MINERVA Internet Libraries have been for the most part hosted by the Internet Archive, with some redundancy of disk backup at the LC. The MINERVA prototype stored HTTrack mirrors of sites, but since their partnership with the Internet Archive, they have used the Alexa .arc format. An analysis of file formats collected shows that for Election 2000 roughly 92% of objects fall into the HTML/TEXT/RTF category, with 7% images and 1% PDF. For the September 11th Archive 82% of objects were HTML/TEXT/RDF, 15% were images, and 1% PDF. For Olympics 2002 85% were HTML/TEXT/RTF, 10% images and 1% PDF. A full reckoning can be found at:
http://www.archimuse.com/mw2003/papers/grotke/grotke.html

Metadata

The Library of Congress has been one of the leaders in promulgating metadata schemes to accommodate preservation and provenance metadata, in addition to descriptive metadata for discovery, to ensure that these assets will be managed far into the future. From the outset a MARC record has been created for each Web site captured and entered into the LC OPAC. With the advent of XML-based encapsulation of metadata such as METS and MODS the LC has adapted its metadata capture to accommodate these innovations. They have contracted with WebArchivist.org to produce MODS records for the sites in the Olympics 2002, Election 2000, Election 2002 and portions of the September 11th Archive. They have also brought in the NDMSO to oversee the production of METS records for the 107th Congress Archive.

For an examination of how METS is particularly poised to accommodate the description and management of archived Web sites, see Appendix 21. The skeleton for a METS object for a Web site can be found in Appendix 22, and a complete METS document for a very simple Web site from the Nigerian Election crawl [*available on request*].

Internet Archive Wayback Machine Profile

The stated mission of the Internet Archive is is to archive at least great portions of the entire Web using donated Alexa content ; to that end they have collected over 300 Terabytes of compressed data since their inception in 1996, adding around 12 T of data to their collection each month. A typical Alexa crawl takes around 8 weeks to complete. Data is stored on hundreds of slightly modified x86 servers running Linux. Each computer has 512Mb of memory and can hold over 1 Terabyte of data on ATA disks. The archival format is the gzipped Alexa .arc file, a 100 MB aggregate of captured files with accompanying server-delivered HTTP headers, along with a .dat file of filtered metadata from HTTP headers and from the achived files themselves; a byte-offset based index accompanies each .arc + .dat Submission Information Package. The data stream isn't altered, but the files are no longer discreet units. URLs are altered to maintain internal consistency and temporal integrity and some javascript and comments are added to the document source on-the-fly upon retrieval.

Preservation/Migration/Emulation

Maintaining copies of the Archive's collections at multiple sites (a mirror is at the modern library of Alexandria in Egypt): part of the collection is already handled this way, and we are proceeding as quickly as possible to do the same with the rest. Although DLT tape is rated to last 30 years, the industry rule of thumb is to migrate data every 10 years. Given developments in computer hardware, we will likely migrate more often than that. As advances are made in software applications, many data formats become obsolete. We will be collecting software and emulators that will aid future researchers, historians, and scholars in their research.

Metadata

"Each ARC file has a corresponding DAT file. The DAT files contain meta-information about each document; outward links that the document contains, the document file format, the document size, etc. Each host provides an index, complete.cdx, located in /0/tmp/. This index may be joined against path_index.txt, located in the same directory, for the full path of the ARC file containing the archived document. In addition to the indices located on each host, the archive also contains an archive-wide index split across 6 remote hosts. These are aliased as index1 - index6. The CDX file on each of these hosts is located in /0/wayback.cdx.gz and is formatted slightly differently than the other CDX files located on each remote host. Refer to the legend on the first line of any CDX file for information on how to interpret the data.**"**

Access Mechanisms

Wayback Machine. "At present, the size of our Web collection is such that using it requires programming skills. However, we are hopeful about the development of tools and methods that will give the general public easy and meaningful access to our collective history. In addition to developing our own collections, we are working to promote the formation of other Internet libraries in the United States and elsewhere."

Administrative Access

Users can apply for *researcher accounts*, which give them access to the files stored files. Unix tools are made available for working with the files.

APPENDIX 14
Evaluation of Prototypes: PANDORA and Kulturarw3
Leslie Myrick

This evaluation takes as its framework the elements of the OAIS Functional Model, addressing issues of Ingest, Archival Storage, Data Management, Administration and Access as they might be applied to a system whose purpose is to harvest a preselected list of political Websites, assign metadata, manage long term preservation and provide access to a designated community through the creation of information packages.

The National Libraries of Australia and Sweden have initiated two very different systems designed to archive digital materials for legal deposit. The former uses a combination of push and pull technology and negotiates relationships with every publisher whose limited number of works have been preselected by a curatorial selector. The latter uses strictly pull processing in a broad swath harvest of what it has designated the Swedish Web without prior negotiation with the myriad of publishers whose works it collects. Both projects are dedicated to the long term preservation of national digital assets, not simply the bytes but the original look and feel of the original object.

The PANDORA Project is part of a larger digital initiative at the NLA, the Digital Services Project, initiated in 1998. They have created an architecture to manage both digitized and born digital content, comprised of five main components: the Digital Object Storage System; the Digital Archiving System; the Digital Collections Manager; a Metadata Repository and Search System; and a Persistent Identifier Resolver Service. The Kulturarw^3 Project is not quite as tightly sutured into the infrastructure of the National Library of Sweden; for instance, there is no expenditure at this time for cataloguing captured materials into the library's OPAC system.

1. Selection/Harvesting Model

The National Library of Australia's PANDORA project and the National Library of Sweden's Kulturarw^3 project have become archetypes of two diametrically opposed approaches to harvesting online digital publications for legal deposit. PANDORA's harvest is selective, concentrating on a predetermined list of electronic publications, while Kulturarw^3 undertakes automated broad swath crawling of the Swedish Web. The NLA has a negotiated relationship with each of its publishers, obviating some of the problems associated with harvesting the deep Web, while the Swedish Library in general does not negotiate in advance, but does have contact with the publishers of Swedish online newspapers, whose sites are harvested daily or weekly and thus visited more frequently by the harvester.

Scope
PANDORA's mandate is to collect scholarly publications and publications of national interest of current and long term research value. Other material may be included as part of a cultural snapshot. In general the material should be relevant to Australia and written by an Australian author, or written by an Australian of recognized authority on a topic of international significance.

The Kulturarw^3 Project has determined the boundaries of the Swedish Web through research into DNS registries. Roughly 45% of their harvest captures the .se domain, 42.5% are .com, .net, .org, .edu registered in Sweden; 12% were in the .nu domain; .05% were *suecana extreana* (externally produced sites about Sweden); 1.2% were IP addresses. Krister Persson comments: The IP addresses mentioned above might well be under the .se (or .org, .com etc.) top level domains. However the links found while collecting links out there have given back these IP addresses.

Size
In 2001, the National Library of Australia was collecting 1250 titles of Australian provenance, following selection guidelines that fit the Library's overall collection development policy. By early in 2003 they were collecting 3287 titles. At that time they had archived 5 million files in 670,000 directories, commanding 134 gigabytes of storage; in 2003 that figure was up to 400 gigabytes comprised of 14 million files. In 2001 the Library of Sweden captured some 30 million files in 1132 gigabytes of storage. The 2002 figures for Kulturarw^3 show 49 million files taking up 1809 GB of disk. Their total to date is 185,95 million files in 5,571 gigabytes.

Coverage

The PANDORA project archives electronic publications with varying rates of change from monographs, with fixed content for the most part, although some evolve over time; journals whose issues appear sequentially and remain fixed as well as some whose contents change over time; and newspapers, which are sampled in snapshots; there are also some digital ephemera that do not have print equivalents, e.g. organizational and personal sites.

The Swedish Library harvests everything from the surface Web that may be reached from ordinary html <A HREF> links. Deep Web content, such as pages containing forms interfacing with database-driven Websites, is not available. What they do derive from database driven sites are static instances of dynamically created pages. In future they will also address problem pages e.g. following pdf links, and XML.

Hardware and Software

In 2001 the NLA was using a Sun E450 and two Sun E250s with a number of Sun Ultra5 workstations. Their database is Oracle. In general the PANDORA project has depended on major investments in proprietary software systems, e.g. the TeraText Content Management System and Oracle database RDBMS. However, they produced the java-based PANDAS system in house using WebObjects. The HTTrack harvester is freely available. In view of persistent clamoring at the gates for open sourcing of PANDAS, PANDORA is considering offering evaluations of PANDAS software as a preliminary to making the source available for a nominal setup/support fee.

The Swedish Library was using a Sun Solaris 450 for harvesting and a 4500 for storage/archiving, but are finding that they need to upgrade the latter. They use Sun workstations for processing and interfacing to the servers. There is no database or content management system. They use an HSM (SAM-FS) to administer the archive with an AML/J tape robot for mass storage and a 1.5 TB disk array as disk cache.

Personnel

The PANDORA project is made up of a manager and five staff members of the Electronic Texts Unit, with significant help from the IT Division, and support from the Preservation Services Branch.

Kulturarw[3] employs a systems manager, two fulltime programmers with occasional input from another programmer.

2. Harvest

Some of the most pertinent ingest issues in harvesting are: 1) the determination of a harvesting model: selective or broad swath? 2) finding or building an appropriate harvester application 3) how to deal with deep Web content that cannot be automatically harvested e.g. database-driven Websites; in some cases dynamic html Web pages generated from .asp, .php, .cgi, .jsp, or .xml; or sites controlled by authentication.

Crawling involves the traversal of a site's tree of links, with parameters set to control how far along the tree to traverse, since hypothetically, an unparameterized crawl could run until it had collected the entire surface Web (should it avoid getting caught in infinite loops from various traps and black holes along the way).

Harvesting the Web is done most effectively by crawlers harnessed to databases into which metadata is extracted, having application classes, database interfaces or daemons that would control scheduling modules, the harvester's activities, link parsers, link filters, indexers and archivers.

The NEDLIB harvester, for instance, deploys interrelated daemons for each of the aforementioned functions, and a MySQL database.
http://www.csc.fi/sovellus/nedlib/ver122/documentation122.doc

The PANDORA project's PANDAS interface incorporates these functionalities as well as editing modules, using Oracle as its backend.
http://pandora.nla.gov.au/manual/pandas/

In large-scale National Library harvests of digital resources for legal deposit both push (publishers' deposit of files via FTP, Web_DAV, or portable media) and pull methods (crawling) are used in various combinations.

Crawler

In the early stages of the PANDORA project, the NLA used a version of the Harvest indexer along with WebZip, but is now deploying HTTrack.
http://www.httrack.com/index.php

Kulturarw^3 uses a much-altered version of the Combine Harvester, an open source application written in Perl for harvesting and threshing (indexing) Web pages developed for use in the DESIRE project and further developed by Netlab for use in Nordic NWI services.
http://www.lub.lu.se/combine/

Crawling Frequency

For the PANDORA project crawling frequency varies according to the resource, and is decided at the point of selection of each resource. Crawls are scheduled according to the inherent dynamic of the title (infrequently for monographs; monthly, quarterly for journals), but special crawls can be initiated as needed. Because publishers sometimes push content to the library, they tend to make notifications of impending changes in format or change of publishing schedule. One-off capture is appropriate for some items (e.g. ephemera).

Kulturarw^3 has done eight complete sweeps of the internet: two in 1997; three in 1998; one in 1999; one in 2000 and one in 2001. A second sweep in 2001 had to be aborted due to a complaint questioning the legality of the harvest (now resolved in their favor). They were 49 million files into their tenth sweep as of Feb 13th. Their recent move to collect newspapers means small but frequent sweeps of each of these sites, sometimes daily.

Distributed Harvest

The PANDORA project depends upon partner institutions such as the State Library of Victoria and the State Library of South Australia for some of its gathering. It is also closely associated with the Tasmanian "Our Digital Island" project.

3. Archiving and Preservation

The overarching question has to be: what exactly is being archived? And to what extent is fidelity to the original look and feel of a site important? Another preliminary question would be: what is the purpose of the archive? Mere preservation, limited access with some parts dark, or full public access? Is this a distributed archive with an interchange of SIPs and/or DIPs from partners? Or is it a monolithic enterprise that performs all the functions of ingest, storage and provision of access?

The technical issues revolve around criteria for building a trusted digital repository whose goal is preservation for the long term, including storage models; compression issues; change and version control; choice of preservation strategies such as migration, emulation, hard- and software museums and mere refreshing of data; and the assurance of the safety, integrity and authenticity of an item through mechanisms such as MD5 checksums and watermarks.

Archive File Format
PANDORA's PANDAS implementation of HTTrack creates a mirror of the Website as well as logs containing HTTP header information and crawler tracking information that can be mined for preservation metadata. Mirrors are stored on a Unix filesystem. An original archived copy is kept and separate work and display copies are created.

Kulturarw^3 uses a multipart MIME file to hold the collection metadata, the header metadata and the file/object itself, very much like a typical output of GNU wget with headers written into the top of the file. They also store several files together in aggregates, much like the IA.

Data Storage Models
The PANDORA project keeps three sets of files: preservation, working and display files.
The Kulturarw^3 project keeps a multipart MIME file on disk and a tape backup of it.

Data Storage
PANDORA has access to disk storage expandable from 2T - 20T in the NLA's SecureData EMC Clarion FC 4700 Digital Object Storage System; they also deploy an HSM system with 8T of tape storage.
Kulturarw^3 uses an HSM system with 20 TB-capacity storage on DLT 7000 tape complementing disk storage of 1.5 TB.

Data Format Issues
The NLA is not planning to normalize its MIME types at this time, although the NAA does prescribe a set of allowable file formats. Because something approaching 90% of the data belongs to one of the four or five most common MIME types they plan to use preservation strategies including migration and emulation to deal with the odd 10%. More than one format is, however, kept of documents originally harvested or deposited in XML or PDF or Word in order to future-proof their continued survival.

The Swedish Library had over 400 MIME types registered in 2001; it is presently collecting close to 800 MIME types. Roughly 90% of them belong to the five most common file types. They plan to use migration and refreshing, rather than emulation, to preserve them.

Long Term Preservation Strategies
PANDORA is planning to wield the full range of preservation strategies: migration, emulation, hard- and software museums, or just plain refreshing for data that cannot be otherwise migrated or emulated.

Kulturarw^3 is depending wholly on storage and migration; they do not plan to use emulation or hoard obsolete soft- or hardware.

Fidelity to the Original
One of PANDORA's driving principles is to try to retain the look and feel of the original site. One copy of each title is kept in its original format as an archival copy. Service copies are created and migrated as necessary when changes in software, file format or technology platform occur. Service copies are altered for the sake of functionality or privacy: mailtos, paypal links as well as all external links are disabled. Some unwanted parts are deleted.

Kulturarw^3 shares the philosophical goal of preserving the original surfing experience. Like the Internet Archive, they have implemented temporal as well as spatial search/viewing. What remains to be developed is a full text / free text index of the indexable material. They are addressing that now as part of the NWA consortium: http://nwa.nb.no.

4. Management and Metadata

Overarching management issues are the use and promulgation of standards; and where possible the adoption of open vs. proprietary standards in the operating system, software, markup, and metadata.

Metadata must be standardized to allow interoperability in the case of distributed archiving systems, and it is generally good practice insofar as metadata usually ends up serving as both a management/preservation tool and an access tool.

All PANDORA metadata is processed along various points of the selection/collection/preservation continuum by PANDAS. Collection metadata is automatically harvested from HTTP header files, but each title is described in MARC by a human cataloguer. Any changes to the Website made through human intervention are manually added to digital provenance metadata.

The NLA published an extensive data dictionary as part of a larger Local Data Model in 1997. http://pandora.nla.gov.au/ldmv2.html. A list of recommended metadata elements (1999) for the PANDORA project can be found at http://www.nla.gov.au/preserve/pmeta.html.

Kulturarw^3 depends on the automatic generation of collection metadata from the crawler. They do not enter cataloguing data into their OPAC for each title at this time.

5. Access

Persistent Identifiers

Having made trial of both PURLs and the Handle System, in 2001 the NLA hired a consultant to reassess their needs. Her report can be found here:
http://www.nla.gov.au/initiatives/persistence/PIcontents.html

For the time being none of DOI, PURL or the Handle System has been found adequate insofar as there is no national or international resolver service. They have therefore implemented their own library-wide PI scheme with an internal resolver service. A persistent identifier for items housed in PANDORA would take the following format: **<collection id>-<work identifier>-<archive date>-<publisher's URI>-<generation code>.** This scheme affords uniqueness, granularity and enough intelligence to enable grouping and relating of versions in the absence of structural metadata.

Kulturarw^3 does not deploy a persistent identifier resolving system per se, but is using a 33-character filename with a timestamp as a persistent identifier internally to the system. As part of the Nordic Metadata project in 1998 they were creating unique, location-independent URNs for their archival objects. http://www.lub.lu.se/metadata/URN-help.html

User Access Mechanisms

For discovery through a title's associated descriptive metadata, a MARC record for each title is entered into the NLA OPAC as well as into the National Bibliographic Database. The PANDORA site includes a search engine, but better search functionality is being investigated. The search engine runs against the TeraText content management system, which replaced MetaStar.

Kulturarw^3 is examining access issues. They have tested the Norwegian product FAST and will examine other software as well. They will most likely go with free text search of the entire MIME type file with metadata embedded with the html text, rather than a system that indexes and searches only metadata such as a MARC record or a finding aid.

Administrative Access

One of the principles of the PANDORA archive is immediate access to external and internal users, but this goal is balanced against the fiduciary interests of publishers when necessary. In relation to commercial publications, periods of restriction on access by external users are negotiated with publishers to protect their income from the title. Periods of restriction range from three months to five years. Access to restricted titles is provided to internal users on a single PC on which electronic copying and communication facilities have been disabled. The Library makes every effort to secure permissions over a title once it has been ingested, withdrawing it only for extenuating legal purposes e.g. because an item is banned or involved in a court case.

Four levels of access have been determined: unrestricted publications; partial commercial restriction; full commercial restriction and full restriction.

The access module for the Kulturarw^3 Project is still in production. Because the National Library of Sweden is collecting Web material without negotiated contracts with publishers, their approach to rights issues will most likely resemble that of the Internet Archive, which allows for an owner of captured Web material to opt out of the archive.

6. Conclusion

The National Libraries of Australia and Sweden stand on different ends of a Web-harvesting spectrum in many respects: the former collects and fully catalogues a predetermined set of titles after negotiating with each publisher, while the latter makes a broad swath capture of the Swedish Web without previous negotiation with publishers and does not enter cataloguing information into the library OPAC. Both libraries in their capacity as holders of legal deposit are collecting and preserving a variety of materials published on the Web that are germane to their national interest.

In a recent phone conversation Margaret Phillips pointed out that the major strength of PANDORA is also its major weakness: that it is a selective archive. Because they harvest a predetermined, circumscribed list they can negotiate with every publisher; evaluate every site captured for its usefulness; and catalogue everything they harvest -- every title has a full MARC record entered into the OPAC and the NBD. The selective model also allows them to check every title for completeness. Roughly 40% of the titles need some sort of active intervention to make them functional.

The primary weakness of the selective model is that selectors are making decisions about what researchers will want in the future, basing their selection upon collection building principles for the print model, which may not be appropriate in the digital realm.

Both PANDORA and Kulturarw^3 have disseminated a rich legacy of reports that can be consulted for statistics; examples of business and logical models; positions on general archival practices; as well as approaches to specific issues in ingest, management, administration, preservation and access that can be applied to archiving political communications on the Web.

7. General Sources for the PANDORA/Kulturarw3 Evaluation

National Library of Australia, Digital Archiving and Preservation at the National Library
http://www.nla.gov.au/initiatives/digarch.html

Margaret Phillips, Archiving the Web: The National Collection of Australian Online Publications
http://www.ndl.go.jp/jp/information/data/nla.doc

Pandora Business Process Model
http://pandora.nla.gov.au/bpm.html

Pandora Logical Data Model
http://pandora.nla.gov.au/ldmv2.html

Pandas Manual
http://pandora.nla.gov.au/manual/pandas

Margaret Phillips, Ensuring Long-term Access to Online Publications
http://www.press.umich.edu/jep/04-04/phillips.html

A. Arvidson et al, The Kulturarw^3 Project - The Royal Swedish Web Archiw3e - An example of "complete" collection of Web pages
http://www.ifla.org/IV/ifla66/papers/154-157e.htm

Kulturarw^3 Description: About the Project
http://www.kb.se/kw3/ENG/Description.htm

Kulturarw^3 Statistics
http://www.kb.se/kw3/ENG/Statistics.htm

Kulturarw^3: To Preserve the Swedish World Wide Web
http://bibnum.bnf.fr/ecdl/2001/sweden/sld001.htm

Additional sources of information used in this evaluation:
Telephone Conversation with Margaret Phillips, 2/3 March 2003.
Email correspondence with Krister Persson and Alan Arvidson, 3-11 March, 2003.

APPENDIX 15

Prototype Evaluation: Web Archiving Project (WARP)

Developing History:
A trial project for a three-year period that started in fiscal 2002 by the NDL (National Diet Library), Japan.

The results of these projects will be used as reference to the Legal Deposit System Council.

Purpose:
To preserve information on the Internet in Japan as cultural property for the sake of future generations.

Scope:
The WARP harvests Web sites selected by the NDL. The WARP consists of two collections: the Website Collection and the Online Periodicals Collections. The Website Collection includes Web sites of governmental organizations, governmental agencies, collaborative organizations and research organizations. The definition of an online periodical for the Online Periodical Collections is a continuously published electronic resource with an identical title and consistent publication frequency.

Size:
- Online journals: 559 titles
- Governmental organizations: 6 Web sites
- Collaborative organizations: 40 Web sites

Overall procedure:
- Selecting the resources to be acquired
- Examining the structure of websites
- Negotiating and contracting for acquisition with publishers
- Specifying the unit of the information resources to be collected
- Creating preliminary metadata
- Setting harvesting, re-harvesting conditions (Including setting URL of starting page, and depth of harvesting)
- Trimming for removing the non-essential parts of the information
- Registering the individual object
- Metadata assignment.

Hardware and Software:
Server – consisting of three servers: a main server, archiving server, and web server.
These servers are also used for other electronic library services.
Main server: FUJITSU GP7000S model 650 (Enterprise 6500)
Archiving server and Web server: FUJITSU GP7000S model T1 (Netra T1)

Disk array (Storage system) – HITACHI Technology SANRISE 2800 724 GB
This storage is also used for other electronic library services.

Database software – SearchServer Version 3.7

Personnel:

Coverage:
The WARP focuses more on surface Web resources rather than deep Web resources.

Harvesting Methods:
Using a Web Robot, wget 1.5.3. (http://www.gnu.org/software/wget/wget.html)

Harvesting Frequency:
- Web sites - monthly
- Online Journals – depending on their publication frequency.

Archive File and Storage:
Roughly 700,000 files, about 35GB

Data Storage Models:

Data Storage:

Data Format Issues:
- HTML – 44.2%;
- JPG – 20.6%;
- GIF – 23.9%;
- PDF – 8.4%;
- others – 2.9%

Fidelity to original:
Retain original data structures of the original Web resources to display the archived resources in the same way with the original resources on the Web.

Preservation/ Migration/ Emulation:

Persistent Identifiers:
Metadata have a new URL for the Web archive and the original URL for the original site as identifiers which link to each resource.

Catalog:
Cataloguing manually before harvesting.

Metadata:
In conformity with the Dublin Core Metadata Element Set, "The NDL Metadata Element Set" was issued as the NDL standard for metadata creation in March2001. Prior to the developments in the legal deposit system, the WARP experimentally adopts the standard and assigns metadata to collected websites and online periodicals. The metadata is filed in “Web Materials Catalogue Database”(temporary name). The metadata can be used for retrieval.

"The NDL Metadata Element Set" is based on the Dublin Core Metadata Element Set, and the NDL adopted some original qualifiers that enable mapping to the JAPAN/MARC format.

Metadata elements in the WARP are as follows: title, author, keyword, description, subject, original URL, new URL, ISSN, ISBN, NDC, language, etc.

Access Mechanisms:
The NDL plans to construct a navigation service based on metadata.

Administrative Access:
Depending on contract conditions with publishers, users can access to the contents of the WARP by Internet or by Intranet.

Look and Feel Issue:

Management Issue:
The NDL is trying to set standards such as harvesting conditions and time interval of re-harvesting by making repeated experiments. Archived resources will also be preserved in other formats such as CD-R.

The NDL manages two copies for each resource: one is for preservation, the other one for the public access.

APPENDIX 16

Harvester Evaluation

November 2003

Harvesting Specifications

Having made preliminary studies of other projects' reports on the difficulties encountered in capturing specific types of Web content such as deep Web material, frame-based, heavily javascripted, or Flash-based sites, along with an examination of reports on the efficacy of their own harvesting applications, we compiled the following list of suggested criteria/questions in evaluating crawlers in general and the three crawlers involved in two capture exercises made by our groups: harvesting selected sites pertaining to the Nigerian Election, and the LANIC time-test crawls.

- **System Configuration:** What are the system requirements? Does the harvester use a database as a backend for managing processes? Daemons? What is the API to the system?
- **Configuration/Default Settings:** What can and cannot be configured? What expertise is needed to configure it? What sort of manuals exist?
- **Problem Files:** How do they each deal with problem files e.g. dynamic pages (.asp, .php, .cgi, .jsp, DHTML, servlet-generated material, database-generated content; applets; forms): do they resolve them into html? What happens in the case of framesets, xml/xslt, downloads? Flash? Javascripts? Does the crawler accept all cookies, or can it be configured to accept only selected cookies?
- **Crawl Methods:** Do they allow for snapshots only? Incremental crawls? In incremental crawls are old files archived or overwritten? In the process of incremental replacement of old files, what are the tests for file modification (checksum, last modified date, etag)?
- **Redundancy:** Can the crawler analyze what percentage of repeated snapshots consist of redundant capture? Or ascertain the percentage of files that were not modified?
- **Archiving format:** How are Web sites and their pages archived? Do they mirror the site? Do they collect all the pages into an aggregate file?
- **Links:** Are internal links rewritten to relative links? Are external HREF links left absolute? Are paypal icons and mailtos disabled?
- **Client scripts:** Are forms, counters and active scripts disabled? In the case of javascript rollovers, how is the second image treated?
- **Metadata:** What range of metadata is captured or filtered out of the client/server transaction and/or the page itself? Where is the metadata stored? What sort of information about sites is provided in crawler logs that requires little or no post-processing? Describe each output log and report.
- **Content:** What did each crawler capture, given the expectations embedded in the configuration? How does the content collected illuminate the ephemerality factor – what changed? When? Which crawler(s) captured changes and which didn't?

Guided by these questions, and in the wake of much further study and a bit of empirical data, the following broad answers have emerged:

- A robust harvesting system such as the Internet Archive focused crawler, the Nordic Group's NEDLIB Harvester, or the Mercator Harvester harnesses a Web crawler to one or more databases and a java- or even a perl-based application that might be composed of specific classes or daemons such as a scheduler, harvester, linkparser, DNS resolver, linkfilter, metaparser and archiver to handle complex functions.
- A supplemental ad hoc crawler to provide supplemental crawls for special events as identified by curators could belong to a class of application along the lines of the NLA PANDAS system, where a simple crawler is supplemented by java-based modules for easier processing.

The ideal harvester should not only create a safe archival copy that may be as simple as an aggregate .arc file, but it should facilitate the rematerialization and access to the service version by copying or representing the original directory structure of the Web site in a zipped mirror. In order to avoid linking out to the live site, it should translate absolute links to relative links that reflect the storage path on the storage file system. It should weed out duplicate files – ideally, it should allow for incremental archiving using some combination of Etags/Last Modified server headers/checksums to prevent at least one subset of duplicate files from being harvested. It should switch off mailtos, payment devices, external links (if desired), forms.

A set of recommendations follow the three case studies below:

Case Study 1: The Alexa and IA Focused Crawlers

An Alexa .arc file is essentially an aggregation of captured HTML and associated server-delivered and crawler-generated metadata for each page bundled into a 100 MB archive file. A generic Alexa .arc is entirely promiscuous and random, according to where that particular crawler wandered in that snapshot, and how much could be packed into a 100 MB file. However, the .arc files that the CRL Project received underwent one further level of handling, to collect each of four region's URLs into its own .arc file(s). Each .arc file is complemented by a .dat (metadata) and a .cdx (index) file. The .dat file contains metadata filtered out of the http headers and the page itself, along with a list of links in each page, broken down by type. The .cdx index collects the URL, arc identifier and checksums along with other HTTP header information into a distillate that can be used to point into the .arc file. There is also a complete .cdx index for all the .arcs residing on a single Internet Archive hard disk.

How does the .arc/.dat/.cdx package stack up as an information package that could be used as an AIP in an OAIS-compliant archive? The major weakness of the .arc format that may consign it to serving as a SIP that will be transformed into an AIP and DIP is that it is not expressed as XML. On the other hand, it exists as text and binary code and is thus a simple lowest common denominator of sorts. On the negative side, the .arc format also offers no readily discernible structural metadata for a Web site, although the links originating in a single page are enumerated in the .dat file, and could be extracted to help recreate the link structure of the original site.

One immediate programming challenge lies in this fact that an .arc file does not recursively mirror the structure of a site, but is a flat aggregation of discrete pages. A possible redress is in part available on the researcher site at the IA, which offers a score of open source scripts for accessing data in an .arc file; one of them, bin_search, outputs a list of all the files that match a certain pattern: it could conceivably gather all the files under a root URL such as www.vaciamiento.com. Then a perl script could be written to parse out the structure into a METS structMap, assuming that an .arc contains all the files in the site; or it could be written across all .arcs belonging to the same snapshot.

Gathering adequate technical metadata for images, executables and any other binary files is another area where the .arc and .dat files will surely have to be supplemented by scripts that extract metadata from multimedia or binary headers, or by human cataloguing using other tools. For instance, total image size in bytes is available in the .dat file as well as MIME type, but the resolution and dimensions of the object are not. Where good image headers exist, a perl script could be devised to parse out information such as the bits of strings that can be found lurking in the binary data archived for each resource in the .arc file. IPTC data can be parsed out using a perl module called Image::IPTCInfo. Issues involved in metadata extraction from other multimedia files and executables will continue to exercise digital preservationists and will be the subject of further study.

In the positive register is the IA .dat file's helpful breakdown of parsed links into HREF links and embedded SRCs, respectively, for the two linked items; this distinction will facilitate processing of the <structMap> and <structLink> sub-elements in a METS object created to wrap Web pages' metadata and files. Two types of md5 checksum are recorded that could be used in an integrity check in a MIX <checksum> element. The IA metaparser also converts Microsoft Code Page entities found in titles into HTML character entities, although on rare occasions it strips them out instead.

Case Study 2: The NEDLIB Harvester

The Technical Team had no hands-on experience with this harvester or its output, but a number of Web-archving projects from the Nordic realms have produced done copious documentation of the strengths and weaknesses of this sytem. A particularly useful description of experiences with NEDLIB can be found in the netarkivet.dk final report, for instance: http://www.netarkivet.dk/rap/Webark-final-rapport-2003.pdf.

The NEDLIB harvester, a product of the Networked European Deposit Library Project, is primarily intended for national libraries to collect and store Web documents as part of their legal deposit activities. The software is available in the public domain, and can be downloaded from: http://www.csc.fi/sovellus/nedlib/ver122/harvester122.tar.Z . The components of the system are nine interrelated daemons to handle complex processes such as scheduling, link parsing, link filtering, parsing

out metadata and monitoring Webservers' performance as well as the harvesting system's performance, with a MySQL relational database backend containing nineteen tables. At this time the harvester does not include an access module/search engine, but an access interface is being developed by the Nordic Web Achive (http://nwa.nb.no/). The first round of capture is a full snapshot, followed by incremental updates.

Disk and software requirements include 64-bit Unix with disk; MySQL; gcc; flex; perl5; tar and gzip. The Linux filesize limit of 2G has been found to be problematic; so also MySQL's 2GB field limit. Solaris 2.6+ on the other hand can support a filesize of 1 TB and so may be preferable in cases where resource- and storage-intensive material such as video is being captured on a large scale. Minimum Processing Requirments are 450Mz Pentium III with Linux kernel 2.x; 512MB - 1 GB memory; 3 G diskspace for OS; 3 G for MySQL; 5 G for harvesting activity; 2 G for accessing data in workspace.

A Good Production Configuration for NEDLIB would be 2 x 400 MHz ultraSparc II with Solaris 2.6+ ; 2 G or more memory; 1 TB tape robot for storage of archive files; 20G diskspace for harvesting; 10-20G for the reqdb daemon; 30G for packing of harvested files (/tmp/); 30G for MySQL; 10G for the OS.

For a longer evaluation of NEDLIB see Appendix 17.

Case Study 3. HTTRACK/PANDAS

Underlying the PANDAS system is HTTrack, a highly configurable offline browser that can mirror Web sites. Harvesting can be automated to a minimal degree by a scheduler in the windows GUI that allows the user to set a deferred harvesting time. It can also be set up as a cron job on Unix. There are numerous similarities between HTTrack and the GNU product wget; my suspicion is that the former is wrapped around the latter with the addition of a number of stealth features such as the ability to overlook robots.txt exclusions and to spoof the user-agent type to avoid robot traps.

HTTrack Features

HTTrack embodies a score of features that could be considered vital in a functional archiving crawler. In a nutshell, it recursively mirrors the structure of the site; can mimic a human user and thus follow rules of proper visiting behavior are parameters involving flow control: e.g. the number of concurrent connections; the number of connections per second and the total bandwidth of the crawl; it can deploy filtering; write dynamic files to .html extensions; has proxy support; incrementally updates the archive; can save previous versions; and has copious logs full of capture metadata.

HTTrack can stay on the same address, directory or domain, or move up or down from the seed URL. Mirroring depth as well as breadth of the harvest of external links are configurable, as are maximum sizes for single files, and overall Web site size. Time constraints such as the overall time limit for the crawl, the transfer rate in bytes/second can also be controlled. These parameters serve both to protect the crawler from crawling into oblivion and to disguise the robot's activity.

It can merely scan and collect information about a Web site; save only html files or only non-html files; grab html first and then non-html. In terms of building the archive, it can grab files in a list, without a recursive reconstruction of the site tree, or it can create a mirror; it can mirror Web sites in interactive mode, allowing selection of pages to harvest. Web site archives can be built in the original structure or in a user-defined structure.

For purposes of circumventing slow or faulty servers, timeout and retry rates are configurable. There are three levels of bailout: host abandon can be set to never, after a certain timeout, or when there is a traffic jam.

HTTrack has well-developed link-parsing functionality: it will parse all links and test all URLs, even forbidden, if so configured. The user can choose to keep the original links or create relative or absolute links where the opposite existed. In general it is good practice to use relative links and perhaps even to concatenate them into a persistent ID, lest the live URL be activated instead of the archived one. For the sake of mirror completeness, HTTrack can replace external links by error pages if external links are not to be collected, or it can simply generate 404 pages for dead internal links.

A major source of mineable metadata, the log output is deluxe and highly configurable. Choices are separate hts-ioinfo.txt and hts-log.txt files, or one integrated file, or no log at all. It also creates a directory of logs and .dat files used for updatng the archive that contains an inventory list of files captured; fairly complete header information with Etags and Date Last Modified metadata, a file containing the HTTP headers and the object itself as well as MD5 checksums. All of this capture metadata is highly extractable into a metadata repository.

The major weakness of HTTrack lies in the fact that out of the box it is manually run and has only a rudimentary scheduler and parser train. There is also no database backend to manage processes or store metadata. For a longer evaluation of HTTrack see Appendix 18.

Harvester Recommendations

The Long Term Resource Management Wireframe calls for federated storage of centrally collected and brokered material by a body such as the Internet Archive using focused crawling strategies and smart crawling technologies that will depend upon a robust system overseen by daemons and either with a scalable database backend or a java-based in-memory frontier for management of the crawl and storage of the metadata, supplemented by local crawling for branding of material and to assure the capture of special events. The centralized crawler would be best supplemented by a system such as PANDAS that has wrapped modular functionalities around the HTTrack crawler that could improve upon its scheduling functionality and allow for quality control and cataloguing into a metadata repository. The repository's data could be repurposed using various interfaces to output metadata packages in MARC, MODS, MIX, and METS for both discovery, preservation needs, as well as access and navigation of archived Web site or Webpage objects.

The harvester should perform an initial snapshot (or perhaps yearly snapshots?) followed by incremental, self-deduping harvests. It should archive pages that have been replaced by modified versions. For ease of service-version generation, it should be capable of writing both to an archival aggregate and to a zipped mirror of the site that can be used as a service version.

It should capture not only HTTP headers delivered with the files but should also employ a metaparser to extract or filter out metadata from the archived files themselves. Sources of this metadata could range from creator-generated <meta> tags to the <title> tag in the HTML page header to <alt> tags for links or images. It should ideally write out the extracted metadata into SQL INSERT or UPDATE statements that could be easily loaded into the metadata repository database.

It should recognize frame-based homepages and be able to render and manage frame-based sites. It should rewrite internal links to relative to the storage path of the file on the archiving filesystem, whether it is an archival file system or a service Webserver file system. It should be smart enough to rewrite often problematic creator-generated relative links to reflect the storage path. It should disable unwanted links, such as mailtos and paypal links. It should ideally be able to parse URLS out of Flash applications. It should rewrite dynamic links with file extensions such as .php and .cgi or .xml to html. It should disable forms. It should be capable of reconstructing complex client-scripted pages using the javascript document function.

A number of robust harvesters are freely available open source applications that use free databases such as MySQL or postgresSQL as their backends; thus costs would accrue not from purchase or support plans but from installation/configuration and internal support that would be included in general systems administration duties.

APPENDIX 17

Harvester Case Study: The NEDLIB Harvester

The NEDLIB harvester, a product of the Networked European Deposit Library Project, is primarily intended for use by national libraries in their collection and storage of Web documents as part of their legal deposit activities. The software is available in the public domain, and can be downloaded as a zipped tarball from:
http://www.csc.fi/sovellus/nedlib/ver122/harvester122.tar.Z

The manual for the release of the NEDLIB Harvester dated 21.09.2001 can be found here:
http://www.csc.fi/sovellus/nedlib/ver122/documentation122.doc

The components of the system consist of nine interrelated daemons and nineteen MySQL tables. At this time the harvester does not include an access module/search engine, but an access interface is being developed by the Nordic Web Achive (http://nwa.nb.no/).; as of this writing it has been successfully paired with the FAST indexer to provide access to harvested collections.

Daemons:
The Scheduler grants new requests to harvesters by monitoring both host Web servers' performance and the internal performance of the harvesting system. The Performance Daemon computes weighted means from service times of hosts; this daemon replaces the linear queue of previous model, which had scalability and performance issues.

The Harvester asks for URLs, fetches documents into the workdir directory, or if it is initially unsuccessful, it stores them in the wqueue table. It then adds each new document to the jobqueue table for further handling by the Linkparser, Linkfilter, Metaparser and Archiver daemons. The Linkparser extracts URLs from the harvested document and adds them to the newURLs table. The Linkfilter weeds out duplicates and already harvested URLs and moves the rest into the goodURLs table. A new daemon, the rdfilter, checks all redirected URLs to prevent the robot from wandering out of the allowed domain/server/Webspace. The Doorman daemon selects an optimal set of new URLs in a text file and feeds it into the Scheduler from the request pool in the reqdb, which is managed by the Reqloader daemon.

The Metaparser uses a perl script to read documents taken from the jobqueue table and parses out metadata into metadata files that sit in the workdir.

The Archiver daemon transfers fetched documents from the workdir to the day directory, creating a separate subdirectory for each collection date. It also creates a URN from the MD5.
In the most recent version there is no longer a choice between full or incremental harvest. The first round must be full, followed by incremental updates.

MySQL Tables:
The following relational tables in a MySQL database support the activities of the various daemons:

- **authority table**: storage of authentication information.
- **config table**: storage for more dynamic configuration options than allowed in the 'definitions.h' file; e.g. 'Maxdepth' field limits a depth of searching process to avoid infinite loops.
- **disallowed table**: information on disallowed URLs, from robots.txt exclusion files.
- **documents table**: information about archived documents. Used for internal purposes only.
- **domains table**: contains all allowed domain suffixes ('.fi', '.csc.fi' etc).
- **goodurls table**: these are new requests passed along by the Linkfilter.
- **hostent table**: a cache for DNS conversion from hostname to IP.

- **hosts table**: allowed and disallowed hosts for this robot.
- **internal wait table**: used for the robot's internal communication between the scheduler and the reqloader.
- **jobqueue table**: the queue for 'harvester-linkparser-metaparser-archiver' pipe, with a limit of 2000 files.
- **knownurls table**: storage for the MD5s of URLs used by the Linkfilter to test for duplicates or already fetched URLS.
- **log table**: list of failed or broken URLs with appropriate error messages.
- **newurls table**: temporary storage for all new URLs from the Linkparser (raw material for the Linkfilter).
- **rdurls table**: table of redirect URLs.
- **robothosts table**: contains the names of those hosts from which 'robot.txt' file has been collected.
- **timespace table**: information about harvesting rounds.
- **urls table**: storage for URLs of collected documents.
- **urlroots table**: seed URLs from which to start harvesting.
- **wqueue table**: This table serves as "waiting room" for URLs that should be revisited because of previous time-out failures, and so on.

The hard- and software requirements for the NEDLIB harvester have been set out on the NEDLIB homepage:

Disk and Software Requirements:
64-bit Unix with disk; MySQL; gcc; flex; perl5; tar and gzip. Linux filesize limit of 2G found to be problematic; so also MySQL's 2GB limit. Solaris 2.6+ can support 1TB filesize.

Minimum Requirments:
450Mz Pentium III with Linux kernel 2.x
512MB - 1 GB memory
3 G diskspace for OS; 3 G for MySQL; 5 G for harvesting activity; 2 G for accessing data in workspace.

A Good Production Configuration for NEDLIB:
2 x 400 MHz ultraSparc II with Solaris 2.6+
2 G or more memory
1 TB tape robot for storage of archive files
20G diskspace for harvesting
10-20G for reqdb
30G for packing of harvested files (/tmp/)
30G for MySQL
10G for OS

Memory, I/O and Storage Issues:

The NEDLIB project has come up with a rule of thumb that 1 TB of storage is needed for 30 million compressed files. Results will, of course, vary according to the nature of the files: HTML pages, Flash applications, image or video or audio files will vary immensely in storage demands.

This harvester is disk-oriented, therefore the major bottleneck will be I/O. Better throughput can be guaranteed by using fibrechannel or SCSI interfaces rather than IDE. Because memory can be an issue with MySQL, it is preferable to run it on a server separate from the harvester itself. This assumes that network can support the I/O. The NEDLIB folks suggest that at least 1G of memory be dedicated to MySQL, and warn very strongly that the default key buffer parameter **must** be increased.

Places where bottlenecks or other problems may occur have been delineated. There are some Unix limitations that will have a bearing on how NEDLIB's harvester performs, e.g. problems with the maximum number of connections -- 1024 is often the default. There is also a limit on the number of files the Unix file system can hold – where one is repeatedly archiving objects that can contain thousands of objects this is certainly a factor. They suggest an inode value of at least one million.

NEDLIB appears to fulfill a good number of the criteria posed by the Tech Team; in the absence of actually using it however, we must base any evaluations we make on other projects' results. One such project is the netarkivet.dk group, whose testbed consists of archived Web sites from the 2001 Danish Elections. They tested an early version of the NEDLIB Harvester along with the open-source crawler wget.

APPENDIX 18

Harvester Case Study: PANDAS/HTTrack

Underlying the PANDAS system is HTTrack, a highly configurable offline browser that can mirror a Web site. Harvesting can be automated to a minimal degree by a scheduler in the Windows GUI that allows the user to set a deferred harvesting time. It can also be set up as a cron job on Unix. There are numerous similarities between HTTrack and the GNU product wget; the former may very well be wrapped around the latter with the addition of a number of useful stealth features such as the ability to overlook robots.txt exclusions and to spoof the user-agent type in order to avoid robot traps.

HTTrack Features

HTTrack embodies a score of features that could be considered vital in a functional archiving crawler. In a nutshell it recursively mirrors the structure of the site; can mimic a human user; can deploy filtering; has proxy support; incrementally updates the archive; can save previous versions; and has copious logs full of capture metadata.

How it visits a site is imminently configurable: it can merely scan and collect information about a Web site using the HEAD HTTP protocol; it can save only html files or only non-html files; it can grab html first and then non-html. In terms of building the archive, it can download a Web site as a flat series of files, without a recursive reconstruction of the site tree, or it can create a mirror; mirroring can be done in interactive mode, with the harvester allowing the human user to select which pages to harvest and which to skip. Web site archives can thus be built in the original structure or in a user-defined structure.

In configuring selection parameters HTTrack can stay on the same address, directory or domain, or move up or down from the seed URL. Mirroring depth as well as breadth of the harvest of external links are configurable, as are maximum sizes for single files, and the overall size limit to be collected. Time constraints such as the overall time limit for the crawl, the transfer rate in bytes/second can also be controlled. These parameters serve both to protect the crawler from crawling into oblivion and to disguise the robot's activity. Also important in making a crawler mimic a human user and thus follow rules of proper visiting behavior are parameters involving flow control: e.g. the number of concurrent connections; the number of connections per second and the total bandwidth of the crawl.

For purposes of circumventing slow or faulty servers, timeout and retry rates are configurable. There are three levels of bailout: host abandon can be set to never, after a certain timeout, or when there is a traffic jam.

Crawling involves the transversal of links, therefore link parsing options are important. HTTrack will parse all links and test all URLs, even forbidden, if so configured. The user can choose to keep the original links or create relative or absolute links where the opposite existed. In general it is good practice to use relative links and perhaps even to concatenate them into a persistent ID, lest the live URL be activated instead of the archived one. For the sake of mirror completeness, HTTrack can replace external links by error pages if external links are not to be collected, or it can simply generate 404 pages for dead internal links.

There are a number of attractive advanced spider options: HTTrack will accept cookies or not; check the doctype if unknown; and parse java classes if desired. It will follow robots.txt and meta tags or not and can spoof a user-agent type.

The log output is deluxe and highly configurable. Choices are separate hts-ioinfo.txt and hts-log.txt files, or one integrated file, or no log at all. HTTrack also creates a directory of logs and .dat files used for incremental updating, that contains lists of files; complete header information with Etags and Last Modified Date metadata for images and pages respectively, a file containing the HTTP headers and the object itself (= 2/3 of an IA .arc file, or a wget log file with embedded headers), as well as MD5 checksums. All of this capture metadata is highly minable. HTTrack can also debug HTTP headers in the logfile. The only failing is the lack of complete HTTP headers.

The major weakness of HTTrack lies in the fact that it must be run manually or by using cron scripts; it has only a rudimentary scheduler and parser train. It cannot be called a robust harvesting system insofar as there are no scheduling or parsing daemons harnessed to a database. To address some of these issues the PANDORA group at the NLA commissioned a programmer in their IT department to write a series of

java modules to wrap around HTTrack. Most of the added functionality serves the quality control, cataloguing and metadata entry needs of the librarians who work in the PANDORA project. The CRL Tech Team was been granted permission to test-drive PANDAS though release date did not occur before the project's end phase.

APPENDIX 19

Summary of Mercator crawl problems

During the investigation, several political and event-based sites were crawled, utilizing a variety of harvesters. The following is a summary of problems surfaced in using Mercator to crawl various sites. This message is drawn from informal discussion and is not a thorough evaluation.

Detailed available crawl data differed in terms of when in the crawl cycle it was collected. For example, the arl, curl and asia crawls were, in each case, the ones I analyzed for mime-type data last year. I am limited in terms of which crawls to use, because other kinds of data (e.g. page counts, etc.) are not available for all crawls, and not all the crawl data is available, because some was lost during the various Mercator crashes.

Several of Peter's crawls used are fairly late crawls (i.e. they occurred many months after crawling for those sites had started, and in the meantime, some sites had gone down, others changed content), while the Nigeria crawl data is an early crawl (i.e. from a few weeks after we had started working with the Nigerian sites). Especially for the more volatile political sites, once a crawl list has been established, basic characterization data is best taken from early crawls, because more and more of the sites become unavailable as time goes on. For example, the Nigerian crawl of May 1, 2003 has good data on 36 of the 37 sites (one site had a robots.txt exclusion). On the other hand, the 9asia crawl was done when about half a dozen of the original sites had already disappeared.

Another problem related to the above concerns the number of sites listed for the mime-type data. I eliminated from the total count any site that had a zero sized mime-counts file. However, the results may also include sites that had gone down and later came up, perhaps with completely different content (e.g. a porn site) because the domain was bought by someone else. These kinds of changes are not well-documented and it is pretty much impossible to determine them retrospectively.

There are similar problems with the page count data. Some sites are shown with an 'x' for the page count. Others show 1 or 2 pages and are most likely not the result of legitimate crawls (Mercator typically shows two pages crawled even for sites it couldn't reach). In crawl 5arl, only 107 out of 130 total sites seem to have produced meaningful page count data.

See Appendix 32 for more details on the Mercator crawl.

APPENDIX 20

Feasibility of Automatic Harvest of Preservation Metadata from Crawler Log Output

Programming modules that permit the capture and filtering out of metadata from the harvesting application and from the Web site data that is collected in the harvesting process should either be built into the harvester itself or built into the system for post-processing. Arguably a full range of metadata from descriptive to technical to structural could conceivably be derived from the material collected. We concentrate here on the descriptive and administrative/technical metadata that we expect to pull out of the process, either into a metadata database repository or into some sort of complex data structure such as a series of hashes of hashes. As we will argue, the complex structure of a Web site in all its temporal versions (and for that matter of any given Web page with its parallel elements of HTML coding, javascript or other client scripting, inline images or video or audio) is best managed by an XML schema such as METS. But in cases where METS implementation is not undertaken, we recommend capturing some metadata about the sum of the parts, including total size, a list of files, the physical file structure, and the link structure of the site.

Capture/Preservation Metadata Desiderata

The technical group in consultation with the curatorial group determined that the following metadata should ideally be captured for the harvest transaction itself, along with metadata for the Web site as a whole and for each individual file:

Host server
- IP address
- Operating System/Web Server Configuration

Harvest/capture transaction
- Timestamp
- Software doing the capture
- Configuration of the software
- HTTP Response Headers:

Status, Content-Length, Content-type, Last Modified Date, Date of transaction
- Errors

Captured Files

(ALL):
- File MIME type
- Filename
- File size
- Last Modified Date
- Creating Software
- Creating Operating System/Hardware
- Checksum/Authenticity (or MM only?)

(HTML):
- Language
- Charset
- Links broken down by type
- META tags: description, keywords HTTP-EQUIV Content-Type
- Embedded scripting
- Encoding

(Multimedia + .exe + binary text downloadables): in flux

For Images: anything that can be extracted from ImageMagick (+ grep against a “strings-ed” version for ColorSpace) and marked up in a MIX extension metadata package.

For .doc, .pdf files: anything for textMD that can be extracted by rtf2txt, pdf2txt, or the Unix strings operator. Many .pdf files, for instance, contain readily harvestable embedded RDF files.

The Web site as a complex digital object
- A summary of files
- The physical file structure
- The link structure

Archive File
- Filename
- Size
- Server/Location

We undertook an evaluation to examine the metadata filtering and logging output from five applications: the IA/Alexa crawler; HTTrack; Mercator; wget and Linklint, with an eye to the range of metadata that is recorded in logs, and to how much might be extracted programmatically into a database or a complex data structure that could output a METS object for each Web site.

Internet Archive/Alexa Crawlers

An Internet Archive SIP consists of three files: the .arc itself, which contains the full text of HTML along with two sets of metadata; the .dat file, a field-value listing of metadata which has been parsed out of the HTML files, e.g. from <meta> tags, along with file offsets, i.e. the physical byte location of the captured resource within the .arc file. Additionally, for each set of arc files there is a .cdx index file with pointer information for all the .arcs in a set or on a hard drive.

The .arc File

Each .arc file contains a filedesc header similar to the following example. A key to the parsed bits follows:

filedesc://IA-001102.arc 0.0.0.0 19960923142103 text/plain 200 - - 0
IA-001102.arc 122

Key:
<URL><sp><IP-address><sp><Archive-date><sp><Content-type><sp><Result-code><sp><Checksum><sp><Location><sp><Offset><sp><Filename><sp><Archive-length>

Next comes a long series of html files preceded by an arc header generated by the crawler and metadata taken from the http headers sent by the host. The arc header takes the following form:

<url><sp><ip-address><sp><archive-date><sp><content-type><sp><result-code><sp><checksum><sp><location><sp><offset><sp><arc filename><sp><length><nl>

and would look like the following:

http://www.dryswamp.edu:80/index.html 127.10.100.2 19961104142103
text/html 200 fac069150613fe55599cc7fa88aa089d - 209 IA-001102.arc 202

The server-delivered http header metadata includes http version; status; date; server; content-type; last modified date; and content-length.

The actual html file follows upon the two headers:

```
<HTML>
<HEAD>HelloWorld</HEAD>
<BODY>
Hello World!!!
</BODY>
</HTML>
```

The metadata that can be extracted from this simple page is rudimentary: a) descriptive: title from the document <head>, last modified date and content type from the http headers; b) structural: none in this case, since there are no links or related files with links; c) preservation: any of the remaining bits of MD delivered by the server or generated by the crawler: IP of the server, status, content-length, identifier of the arc file, location in the arc file, date of the crawl, and so on.

The key to .dat and .cdx files

.dat and .cdx files contain the following letters:

A canonized url
B news group
C rulespace category ***
D compressed dat file offset
F canonized frame
G multi-columm language description (* soon)
H canonized host
I canonized image
J canonized jump point
K Some weird FBIS what's changed kinda thing
L canonized link
M meta tags (AIF) *
N massaged url
P canonized path
Q language string
R canonized redirect
U uniqness ***
V compressed arc file offset *
X canonized url in other href tages
Y canonized url in other src tags
Z canonized url found in script
a original url **
b date **
c old style checksum *
d uncompressed dat file offset
e IP **
f frame *
g file name
h original host
i image *
j original jump point
k new style checksum *
l link *
m mime type of original document *
n arc document length *
o port
p original path
r redirect *
s response code *
t title *
v uncompressed arc file offset *
x url in other href tages *
y url in other src tags *
z url found in script *

* in alexa-made dat file
** in alexa-made dat file meta-data line
*** future data

The .cdx file
The .cdx file contains a line which summarizes each site having the format CDX A b e a m s c k r V v D d g M n. This translates to:

> <url><sp><date><sp><IP><sp><original URL><sp><mime type><sp><response code><sp><old style checksum><sp><new style checksum><sp><redirect><sp><compressed offset><sp><uncompressed offset><sp><compressed dat file offset><sp><uncompressed dat file offset><sp><file name><sp><meta tags><arc document length>

A typical .cdx entry:

> 0-0-0checkmate.com/Bugs/Insect_Habitats.html 20010424210312 209.52.183.152 0-0-0checkmate.com:80/Bugs/Insect_Habitats.html text/html 200 d520038e97d7538855715ddcba613d41 30025030eeb72e9345cc2ddf8b5ff218 - 47392928 145482381 4426829 15345336 DE_crawl3.20010424210104 – 635

The .cdx index is not an index intended for content discovery; it is essentially collects the URL, arc identifier and checksums along with other http header information into a distillate that can be used to point into the .arc file.

HTTrack

HTTrack produces 5 or 6 logs containing metadata for its own use in keeping track of what has already been captured in incremental crawling; but these can be programmatically extracted to serve as preservation metadata in a Web archive.

A. **new.lst** That merely lists the files captured in capture order. It is not particularly useful in itself, but could be used to reconstruct the physical file structure on the server with a recursive script, for instance.

B. **new.txt**
A tabular log showing:

> time size/remotesize flags(update,range, fileresponse, modified, chunked, gzipped) statuscode status (servermsg) MIME etag/date URL localfile (from URL)

<snip>
09:06:36 391/391 ---MC- 404 error ('Not%20Found')text/html date:Sat,%2015%20Mar%202003%2014:17:21%20GMT www.vaciamiento.com/robots.txt (from)
09:06:37 40588/40588 ---MC- 200 added ('OK') text/html date:Sat,%2015%20Mar%202003%2014:17:21%20GMT www.vaciamiento.com/ C:/My%20Web%20Sites/vaciamiento1/www.vaciamiento.com/index.html (from)
09:06:39 811/811 ---M-- 200 added ('OK') image/gif etag:%2266c130-32b-3d6cc2db%22 www.vaciamiento.com/images/tex-salvemosarg.gif C:/My%20Web%20Sites/vaciamiento1/www.vaciamiento.com/images/tex-salvemosarg.gif (from www.vaciamiento.com/)
</snip>

C. **New.dat**
A multipart MIME file containing metadata, a checksum and the content of HTML files.
N. B. HTTrack's logging module has its own ideas about whitespace – any application extracting metadata will have to parse the 200 response code out of 2003, which is apparently a concatenation of two pieces of metadata, for instance.
1) for an image file:

<snip>
2 [unknown]
329a4d859b1ea195895bbe15f061ec26823 [checksum]
2003 [response code = 200, ok]
8112 [content-length 811]
OK9 [server message]
image/gif29 [mime type]
Wed, 28 Aug 2002 12:32:27 GMT21 [last modified date]
"66c130-32b-3d6cc2db"0 [etag]
</snip>

2) for an HTML file:
<snip>
2004 [200 = response code, i.e. successful]
94402 [content-length sent 9440]
OK9 [server message]
text/html29 [file type]
Sat, 15 Mar 2003 14:22:43 GMT0 [date]
0 [unknown]
0 [unknown]
3
HTS4
9440[content-length received]
<html>
<head>
[N.B. These meta tags can be mined for additional metadaa, e.g. charset and the editor that created the Web page]
<meta http-equiv="Content-Type" content="text/html; charset=windows-1252">
<meta name="GENERATOR" content="Microsoft FrontPage 4.0">
<meta name="ProgId" content="FrontPage.Editor.Document">
<title>Nuevo gabinete de Duhalde</title>
</head>

<body>[. . .]</body>
</html>
2 [unknown]
32ed4f9f49d01db15f48c1a5b19f8744703 [checksum]
</snip>

D. **winprofile.ini** on windows [= **httrack.conf** in unix]

An accounting of the configuration of the crawler, for instance, whether to honor robots.txt, whether to parse java files; how deep and broad to crawl; how many retries; whether to accept cookies, and so on.

This file is generated by the windows version and kept on record for subsequent crawls of the same site. On UNIX it would be manually configured by the user.

<snip>
Near=0
Test=0
ParseAll=1
HTMLFirst=0
Cache=1
NoRecatch=0

Dos=0
Index=0
WordIndex=0
Log=1
RemoveTimeout=0
RemoveRateout=0
FollowRobotsTxt=2
NoErrorPages=0
NoExternalPages=0
NoPwdInPages=0
NoQueryStrings=0
NoPurgeOldFiles=0
Cookies=1
CheckType=1
ParseJava=1
</snip>

Some of the features of the crawl under the configuration given above are that the crawler did not follow near files, did parse java, was not asked to collect the HTML before the binary resources, did not make an index, had logging turned on, accepted cookies, and so on.

E. **hts-info.txt**

This file logs the request and response transactions as they occur.
It contains a handful of HTTP headers including server configuration, when available:

Server: Apache/1.3.26 (Unix) Chili!Soft-ASP/3.6.2 PHP/4.2.2 FrontPage/5.0.2.2510 mod_perl/1.27 mod_ssl/2.8.9 OpenSSL/0.9.6b

A fair amount can be surmised from this information: e.g. it is likely that the site uses dynamic scripting in the form of active server pages and PHP, since ChiliSoft is a significant investment.

The log that follows records the request and response transactions between the client and the server. First the crawler asks to consult the robots.txt page, then moves on to the index page:

<snip>
request for www.vaciamiento.com/robots.txt:
<<< GET /robots.txt HTTP/1.1
<<< Connection: close
<<< Host: www.vaciamiento.com
<<< User-Agent: Mozilla/4.05 [fr] (Win98; I)
<<< Accept: image/gif, image/x-xbitmap, image/jpeg, image/pjpeg, image/svg+xml, */*
<<< Accept-Language: en, *
<<< Accept-Charset: iso-8859-1, *
<<< Accept-Encoding: gzip, deflate, compress, identity

request for www.vaciamiento.com/:
<<< GET / HTTP/1.1
<<< Connection: close
<<< Host: www.vaciamiento.com
<<< User-Agent: Mozilla/4.05 [fr] (Win98; I)
<<< Accept: image/gif, image/x-xbitmap, image/jpeg, image/pjpeg, image/svg+xml, */*
<<< Accept-Language: en, *
<<< Accept-Charset: iso-8859-1, *

<<< Accept-Encoding: gzip, deflate, compress, identity

response for www.vaciamiento.com/robots.txt:
code=404
>>> HTTP/1.1 404 Not Found
>>> Date: Sat, 15 Mar 2003 14:17:21 GMT
>>> Server: Apache/1.3.26 (Unix) Chili!Soft-ASP/3.6.2 PHP/4.2.2 FrontPage/5.0.2.2510
mod_perl/1.27 mod_ssl/2.8.9 OpenSSL/0.9.6b
>>> Connection: close
>>> Transfer-Encoding: chunked
>>> Content-Type: text/html; charset=iso-8859-1
</snip>

F. **new.idx**

This is an index file that consists of a pointer to the file's physical location in new.dat; it is not a discovery index per se.

<snip>
www.vaciamiento.com
/corralito.htm
366659
78
//[HTML-MD5]//
C:/My Web Sites/vaciamiento1/www.vaciamiento.com/corralito.htm
371200
42
</snip>

Mercator

Mercator crawls a site or sites starting from a pool of seed URLs. As documents are downloaded, they are analyzed on the fly by a series of custom analyzer modules. The output of the modules can be piped to other analyzers or can be written to files. The following files and the included metadata are common to all the crawls being done by Cornell.

clock.000000
Information about checkpointing, in case a crawl is interrupted.

config.sx
The editable configuration file for the crawl. Among the attributes that can be changed are the Seed URLs, the Filter strings limiting the crawl, Politeness rules, and Analyzer add-ins for massaging the raw data during run-time.

Example:

```
((("SeedURLs" ("http://afenifere.virtualave.net/"))
 ("Filter" ("Domain" ".virtualave.net"))
 ("AtraxMachines" ("crawler0"))
 ("DnsClass"
  ("mercator.dns.MercatorDNS"
   (("NameServer" "localhost")
    ("AssumeCanonicalHost" "true")
    ("CacheOracleClass"
     ("mercator.cache.ClockReplacementOracle" (("LogSize" "17")))))))
```

```
 ("TimeLimitSecs" "86400")
 ("CheckpointFreq" "96400")
 ("NumberThreads" "80")
 ("MaxDepth" "100")
 ("WorkDirPath" "/misc1/cornell/kehoe/crawls/nigelec.05013/")
 ("StableDirPattern" "'afenifere.virtualave.net'")
 ("FrontierClass"
  ("mercator.frontier.PoliteStaticFrontier"
   (("QueueThreadRatio" "3.0")
    ("QueueClass" ("mercator.queue.BuffDiskQueue" (("PoolSize" "200"))))
    ("PoolBase" "pools-1051815710173")
    ("PolitenessFactor" "10")
    ("StrictPoliteness" "true")
    ("MinRestSecs" "0")
    ("MaxRestSecs" "2147483647"))))
 ("URLSetClass"
  ("mercator.urlset.ContextAwareURLSet"
   (("CanonicalizeHost" "false")
    ("SizeLimit" "-1")
    ("LogURLTrace" "false")
    ("LogBuffSize" "21")
    ("LogSpineSize" "16")
    ("LogCacheSize" "18"))))
 ("Protocols"
  (("ftp"
    ("mercator.protocol.ftp.FTPProtocol"
     (("EmailAddress" "wrk1@cornell.edu")
      ("AllowNonASCIIinURLs" "false"))))
("http"
    ("mercator.protocol.http.ContextAwareMercatorProtocol"
     (("UserAgentMailbox" "wrk1@cornell.edu")
      ("UserAgentBase" "Mercator")
      ("UserAgentVersion" "1.0")
      ("ProxyRules" ())
      ("SocketTimeoutSecs" "60")
      ("NumberTries" "3")
      ("AllowNonASCIIinURLs" "false")
      ("AcceptHeader" "null")
      ("ConditionalFetch" "false"))))))
 ("Analyzers"
  (("text/html"
    ("mercator.analyzer.html.PageData" ())
    ("mercator.analyzer.html.CountingLinkExtractor" ())
    ("mercator.analyzer.SaveDoc" ())
    ("mercator.analyzer.html.HttpHeaderExtractor" ()))
   ("image/gif" ("mercator.analyzer.gif.GifHistograms" ()))
   ("ftp/directory" ("mercator.analyzer.ftpdir.CountingLinkExtractor" ()))))
 ("ExitOnEmpty" "true")
 ("ContinuousCrawl" "false")
 ("CheckpointVersion" "0")
 ("StartDate" "1051815709094")
 ("LogFiles" ("stdout" "crawl-log.txt"))
 ("LogRobotsCacheSize" "17")
 ("FilterTwice" "false")
 ("DocFPSetClass"
```

```
 ("mercator.fpset.DiskFPSet3"
  (("LogCacheSize" "-1")
   ("LogBuffSz" "18")
   ("LogBuffTblLoad" "4")
   ("BucketSizeIncr" "4"))))
("PerDocLoggerClass" ("mercator.core.FilePerDocLogger" ()))
("LogRISMemSize" "16")
("LogRISMaxSize" "20"))
```

crawl-log.txt

A log of the actions of the crawl. It also includes the config.sx file. That section is omitted from the following example:

```
Start Date: Thu May 01 12:02:33 PDT 2003
Host: crawler0-complaints-to-admin.webresearch.pa-x.dec.com
System properties:
   java.runtime.name = Java(TM) 2 Runtime Environment, Standard Edition
   java.runtime.version = 1.3.1-beta2
   java.vendor = Compaq Computer Corp.
   java.version = 1.3.1
   java.vm.info = native threads, mixed mode, 07/31/2001-09:18
   java.vm.name = Fast VM
   java.vm.vendor = Compaq Computer Corp.
   java.vm.version = 1.3.1-beta2
   os.arch = alpha
   os.name = OSF1
   os.version = V5.1
Overridden crawler attributes:
(("SeedURLs" ("http://buhariokadigbo.com/"))
 ("Filter" ("Domain" "buhariokadigbo.com"))
[…omitted configuration…]

   Elapsed Discovered   Frontier Downloaded     Unique Overall Current Overall Current  DownLoad
Wait  Memory
      Time      Pages      Pages      Pages      Pages  docs/s  docs/s  KB/sec  KB/sec  Failures Thds
Usage
----------- ---------- ---------- ---------- ---------- ------- ------- ------- ------- --------- ---- -------
Workers started.
:     10.1        53        29        25        24    2.47    2.47    19.6    19.6         0   78   68699
:     20.1        54        11        44        43    2.19    1.90    22.7    25.7         0   78   86148
:     30.1        54         1        54        53    1.79    1.00    18.6    10.6         0   79   92178
:     40.1        54         1        54        53    1.35    0.00    14.0     0.0         0   79   92178
Flushing 54 entries to /misc1/cornell/kehoe/crawls/nigelec.05013/buhariokadigbo.com/urlfpset.curr
took 5 ms
:     50.2        54         0        56        53    1.12    0.20    11.6     2.2         0   80   92178
Frontier is empty -- terminating crawl
Stopping workers...
Workers stopped.
Flushing DiskFPSet to /misc1/cornell/kehoe/crawls/nigelec.05013/buhariokadigbo.com/docfpset.curr
took 16 ms
  contains = 54; hits1 = 0; hits2 = 1; hits3 = 0
  diskLookups = 53; seeks = 0; reads = 0
Crawl terminated at Thu May 01 12:03:25 PDT 2003
```

Docs-null.0000.000000

This file contains the documents as received by the client, including HTTP headers and the HTML document. Multiple pages are aggregated into one or more files. The pages are delimited by a leading line in the form “-----size-in-bytes-----URL”

Example:

```
-----22127-----http://buhariokadigbo.com:80/
HTTP/1.1 200 OK
Date: Thu, 01 May 2003 18:02:35 GMT
Server: Apache/1.3.27 (Unix) mod_throttle/3.1.2 PHP/4.3.0 mod_ssl/2.8.11 OpenSSL/0.9.6g
FrontPage/5.0.2.2510
Last-Modified: Sat, 12 Apr 2003 12:59:44 GMT
ETag: "3f80b-5525-3e980dc0"
Accept-Ranges: bytes
Content-Length: 21797
Connection: close
Content-Type: text/html

<html>

<head>
<meta http-equiv="Content-Type" content="text/html; charset=windows-1252">
<meta http-equiv="Content-Language" content="en-us">


<title>Buhari-Okadigbo Home Page</title>
<meta name="GENERATOR" content="Microsoft FrontPage 4.0">
[SNIP]
```

gif-stats.logA summary of some information about the gif files encountered. This is the output of the "mercator.analyzer.gif.GifHistograms" run-time analyzer.

http-status-histo.log
A summary of HTTP status codes for all the downloaded pages.

Example:

```
============ Statistics for Checkpoint 0 ============

-9999 = download failures / no status code
-9998 = download disallowed by /robots.txt file

Total elements: 54
Aggregate stats: min = 200, mean = 200.00, max = 200
200 --> 54 (100.0%)
```

httpHeaders-0000.000000

just the HTTP headers for each page, delimited by a leading line in the form “-----size-in-bytes-----URL”. The output of the “mercator.analyzer.html.HttpHeaderExtractor” run-time analyzer.

Example:

```
-----22127-----http://buhariokadigbo.com:80/
HTTP/1.1 200 OK
Date: Thu, 01 May 2003 18:02:35 GMT
Server: Apache/1.3.27 (Unix) mod_throttle/3.1.2 PHP/4.3.0 mod_ssl/2.8.11 OpenSSL/0.9.6g
FrontPage/5.0.2.2510
Last-Modified: Sat, 12 Apr 2003 12:59:44 GMT
ETag: "3f80b-5525-3e980dc0"
Accept-Ranges: bytes
Content-Length: 21797
Connection: close
Content-Type: text/html
-----6945-----http://buhariokadigbo.com:80/Main/feedback.htm
HTTP/1.1 200 OK
Date: Thu, 01 May 2003 18:02:37 GMT
Server: Apache/1.3.27 (Unix) mod_throttle/3.1.2 PHP/4.3.0 mod_ssl/2.8.11 OpenSSL/0.9.6g
FrontPage/5.0.2.2510
Last-Modified: Fri, 11 Apr 2003 01:09:47 GMT
ETag: "edba8-19d8-3e9615db"
Accept-Ranges: bytes
Content-Length: 6616
Connection: close
Content-Type: text/html
```

mime-counts.000000

A binary file containing the number of files downloaded by mime type.

pageData.000000

A set of files–the output of the mercator.analyzer.html.PageData run-time analyzer–containing an XML-encoded listing of the components of the page.
The elements:

- <URL>--the page's locator
- <PAGEHEADER>--the content of the page <head></head> element
- <JAPPLETCOUNT>--the number of Applets included in the page
- <FORMCOUNT>--the number of forms on the page
- <JSCRIPTCOUNT>--the number of javascripts on the page
- <ILINKS>--the number of links from this page to other pages within the site.
- <ILINK>--the actual internal links
- <ELINKS>--the number of links from this page to external pages.

Example:

```
<?xml version="1.0" encoding="UTF-8"?>
<PAGEPROFILE>
<URL>http://buhariokadigbo.com:80/Press%20Releases/Buhari-
Okadigbo%20Supporters%20in%20United%20States%20Urge%20Nigerian%20Vo\
ters.htm</URL>
<PAGEHEADER>
<head>
<meta content="HTML Tidy, see www.w3.org" name="generator"/>
<meta content="text/html; charset=windows-1252" http-equiv="Content-Type"/>
<meta content="Microsoft FrontPage 4.0" name="GENERATOR"/>
<meta content="FrontPage.Editor.Document" name="ProgId"/>
<title>Buhari/Okadigbo Supporters in United States Urge Nigerian Voters</title>
<meta content="copy-of-straight-edge 000, default" name="Microsoft Theme"/>
```

```
<meta content="tlb, default" name="Microsoft Border"/></head>
</PAGEHEADER>
<JAPPLETCOUNT>0</JAPPLETCOUNT>
<FORMCOUNT>0</FORMCOUNT>
<JSCRIPTCOUNT>0</JSCRIPTCOUNT>
<ILINKS count ="5">
        <ILINK>../images/map_nigeria.gif</ILINK>
        <ILINK>../Links.htm</ILINK>
        <ILINK>mailto:info@buhariokadigbo.com</ILINK>
        <ILINK>mailto:cakukwe@att.net</ILINK>
        <ILINK>mailto:Webmaster@buhariOkadigbo.com</ILINK>
</ILINKS>
<ELINKS count ="1">
        <ELINK> http://www.buhariokadigbo.com/</ELINK>
</ELINKS>
</PAGEPROFILE>
```

robotsLog.000000
The URLs and HTTP status codes for attempted downloads of presumed robots.txt files. In the following example, Mercator tried the URL, but got a “Page not found” return.

Example:

```
http://buhariokadigbo.com:80/robots.txt
   Status code 404
```

timings.000000
Mostly more metadata about the crawl. The first line of the example contains metadata about the crawl. The second contains metadata about the page, including HTTP status code, length in bytes, a checksum for the page, and the URL.

Example:

```
start time  ID  blocked  RW lock    fetch  process      dns    total
----------- --- -------- -------- -------- -------- -------- --------
        16   0        2        1      710      550       47     1310

code  length      fingerprint   URL
----- -------- ---------------- ------------------------------------
200   21797    1610bc292868e162 http://buhariokadigbo.com:80/
```

Wget

So-called verbose logging from wget records capture time; captured URL and the local URL as it is mirrored on the file system; connection status; content-length and content-type; and download connection speed. It is essentially a record of the request and response history with a minimum of http headers incorporated.

```
<snip>
--10:54:20--  http://dlib.nyu.edu/webarchive
          => `dlib.nyu.edu/webarchive'
Connecting to dlib.nyu.edu:80... connected!
HTTP request sent, awaiting response... 301 Moved Permanently
Location: http://dlib.nyu.edu/webarchive/ [following]
--10:54:20--  http://dlib.nyu.edu/webarchive/
```

```
        => `dlib.nyu.edu/webarchive/index.html'
    Connecting to dlib.nyu.edu:80... connected!
    HTTP request sent, awaiting response... 200 OK
    Length: 3,592 [text/html]

      0K -> ...                                 [100%]

    10:54:20 (3.43 MB/s) - `dlib.nyu.edu/webarchive/index.html' saved [3592/3592]

    Loading robots.txt; please ignore errors.
    --10:54:21--  http://dlib.nyu.edu/robots.txt
        => `dlib.nyu.edu/robots.txt'
    Connecting to dlib.nyu.edu:80... connected!
    HTTP request sent, awaiting response... 200 OK
    Length: 168 [text/plain]

      0K ->                                     [100%]
</snip>
    10:54:21 (164.06 KB/s) - `dlib.nyu.edu/robots.txt' saved [168/168]

    --10:54:21--  http://dlib.nyu.edu/webarchive/prototypes.html
        => `dlib.nyu.edu/webarchive/prototypes.html'
    Connecting to dlib.nyu.edu:80... connected!
    HTTP request sent, awaiting response... 200 OK
    Length: 4,308 [text/html]

      0K -> ....                                [100%]

    10:54:21 (4.11 MB/s) - `dlib.nyu.edu/webarchive/prototypes.html' saved [4308/4308]

    --10:54:21--  http://dlib.nyu.edu/webarchive/index.html
        => `dlib.nyu.edu/webarchive/index.html'
    Connecting to dlib.nyu.edu:80... connected!
    HTTP request sent, awaiting response... 200 OK
    Length: 3,592 [text/html]

      0K -> ...                                 [100%]
</snip>
```

Linklint

Linklint is essentially a link checker and not a mirroring crawler per se, but it can crawl a site and produce a panoply of useful reports that monitor the state of the link structure, as well as giving a summary of the files contained in the site.

The logs for a single HTTrack capture of one of the Nigerian Election sites can be found here: http://dlib.nyu.edu/webarchive/linklintlogs/socialistnigeria/index.html

A. **summary.txt**
This log reports on the overall structure and health of the site – the total number of directories; whether there is a default index; the total number of files and its breakdown between HTML, image and other; external links e.g. mailtos that should be disabled; missing links, missing internal anchors.

It records the root of the site; date of capture and software version at the top of the file, and specifies errors at the bottom.

<snip>
file: summary.txt
root: /www.socialistnigeria.org
date: Sun, 27 Apr 2003 11:43:17 (local)
Linklint version: 2.3.5

Linklint found 102 files in 16 directories and checked 91 html files.
There were no missing files. No files had broken links.
1 error, no warnings.

found 16 directories with files
found   1 default index
found 90 html files
found   2 image files
found   9 other files
found   1 http link
found   3 mailto links
found   6 named anchors
----- 1 action skipped
ERROR   1 missing named anchor
</snip>

B. **log.txt**
Records the progress of the crawl.

C. **dir.txt**
Gives a list of directories

D. **file.txt**
Lists files by type
http://dlib.nyu.edu/webarchive/linklintlogs/socialistnigeria/file.htm

E. **fileX.txt**
Records cross-referencing/cross-linking amongst the files
http://dlib.nyu.edu/webarchive/linklintlogs/socialistnigeria/fileX.htm

F. **fileF.txt**
Records the link structure of the Web site by recording the forward links found in every file.
http://dlib.nyu.edu/webarchive/linklintlogs/socialistnigeria/fileF.htm

G. **remote.txt** and remoteX.txt
Record the external http links found and the files that contain them.

H. **anchors.txt** and anchorsX.txt
Record named anchors and where they are found.

I. **action.txt** and actionX.txt
Record actions that were skipped and where those action links occurred. This involves form input for the most part.
In the following case the link is to an external site and to a perl script:
http://dlib.nyu.edu/webarchive/linklintlogs/socialistnigeria/action.htm

J. **errorA.txt** and errorAX.txt
Records errors in named anchors and the files in which they occur.

N.B. The following occur in other mirrors:

K. **warn.txt**
Various warnings show up here. In this case no single index.html file was found (since the homepage is a frameset) :
http://dlib.nyu.edu/webarchive/linklintlogs/allidemoUK/warn.htm

L. **imgmap.txt** and imgmapX.txt
Records named image maps and the files in which they occur.

http://dlib.nyu.edu/webarchive/linklintlogs/peoplesmandate/imgmap.htm

Recommendations

A combination of focused IA crawling, Mercator or an application such as PANDAS wrapped around HTTrack paired with Linklint would appear to provide a sufficient amount of capture metadata for preservation. A further application e.g. ImageMagick for images, should be considered to extract more specific preservation metadata e.g. pixels wide and pixels high; bits per sample; compression and so on.

The mechanism for extracting the metadata from these logs could be as simple as a series of perl scripts. Any scripting language that can do system calls, recurse through directories and files, read in files in binmode, and process text will suffice for this task. These scripts would output SQL statements or CSV load files that can be dumped into a database for further processing. In progress is a series of perl scripts that can make a series of SQL queries against a Web site's metadata in the database and create the dmdSec, amdSec, fileSec, structMap and structLink for METS description/encapsulation of a site.