IIPC Preservation Working Group Table of Threats and Potential Solutions
Initially prepared as a result of an internal NLA Workshop on 5th June, amended as a result of teleconferences held with PWG members and additional feedback and discussion.

Table of Contents

Note: The following threats have been removed from current discussion as they offer little scope for practical action at this stage. They may however be added as a general list of threats later if the work of the PWG is extended:

- Media Obsolescence [originally Threat 5]
- Unable to recreate 'the experience' [originally Threat 14]
- Version control [originally Threat 15]
- Unable to assess success [originally Threat 16]

Threat/consequence	**Potential Standards/Tools/Approaches**	**Other Comments**
1. Not taking action [loss of material]	OAIS	OAIS provides a structured approach and enables organisations to proceed incrementally towards a fully operational repository.
	TRAC	TRAC provides a checklist, some of which is applicable to web archives. TRAC should be analysed by the sub-group on organisational issues for its relevance to web archives
	DRAMBORA	If any PWG member has experience of using DRAMBORA it would be useful to have some feedback. This could also be fitted under the overall task of assessing available tools.
	Mission Statement that reflects a commitment to digital preservation [**TRAC Ref A1.1**]	This is a first step to making this activity a strategic and organisational priority.
	Business and Risk Management Plans [**TRAC Ref A4.1**]	This will be useful to justify embarking on a web archiving program in the face of competing priorities.
	Harvesting Tools (e.g. Heritrix)	Emphasis to date has been on capturing material so that preservation strategies can be worked on later. See also specific threats associated with ingest – Threats 2, 3, 7,8
	DCC Catalogues of Web Archiving Tools	This is a lengthy document, difficult to find a tool for a specific purpose, and some tools are dated (e.g. cites NWA Toolset, but not WERA). Note: This is a general point applicable to several tools being developed, that they can be quite labour intensive to use so their benefit (particularly with regard to Threat number 12) is therefore reduced.

Threat/consequence	Potential Standards/Tools/Approaches	Other Comments
2. Viruses [data corrupted; archive could become unworkable] Is generally a higher risk for whole domain harvesting than for manually selected archiving (though the latter is of course much more resource intensive) **TRAC Reference** **C3 Security**	Virus Checking	Delays ingest workflow [not really practical for large scale web harvests?]
	Quarantine at ingest	Delays ingest workflow [not really practical for large scale web harvests?]
	Effective Firewalls	For PANDORA, malicious code and embedded spyware is a potential threat as, unless it causes problems as it's brought into the archive, there is no way of checking. However, as spyware requires a viewer to execute, the *preservation* copy within the archive is at low risk, though the *presentation* copy may cause problems for external users.
3. Data Corruption [unable to read data; unable to verify authenticity]	Error checking (e.g. checksums) **TRAC References** **A3.8** Repository commits to defining, collecting, tracking, and providing on demand, its information security measurements **B4.4** Repository actively monitors integrity of archival objects (i.e. AIPs), **C1.5** Repository has effective mechanisms to detect bit corruption or loss. **C1.6** Repository reports to its administration all incidents of data corruption or loss, and steps taken to repair/replace corrupt or lost data.	The NLA Repository (DOSS) takes a checksum (or similar) of each AIP when it is submitted to the repository. This checksum is then used to check the integrity of the object when it is retrieved at some time in the future. This error checking only happens when a request for retrieval is processed. NLA therefore believes it meets TRAC reference C1.5, but not the others cited here.

Threat/consequence	Potential Standards/Tools/Approaches	Other Comments
4. Media Failure **TRAC Reference C1.7** Repository has defined processes for storage media and/or hardware change (e.g. refreshing, migration.)	Select appropriate media	Note: Most NLA media failures are caused by tapes snapping. In these cases, the second copy or backup may be retrieved, as necessary.
	Undertake programme of regular refreshment	
	Multiple back-ups (including redundant back-ups incase one set of media fails. Ideally 3 copies of each instance)	
	Keep in dynamic systems but as hard drives with RAID redundancies and hot spares.	
5. Disaster [loss of data] **TRAC Reference C3.4** Repository has suitable written disaster preparedness and recovery plan(s), including at least one off-site backup of all preserved information together with an off-site copy of the recovery plan(s).	Data Storage standards, e.g RAID	All approaches should work for web archives
	Data back-up and recover regime	
	Disaster prevention and recovery plan	Note: At NLA IT disaster planning is IT's responsibility. Special IT skills would be required to implement a recovery operation.
	Code of practice for information security management: ISO 17799:2005	

Threat/consequence	Potential Standards/Tools/Approaches	Other Comments
6. Inadequate documentation [Not knowing what we have. Unable to establish provenance. Unable to maintain authenticity. Unable to provide access. Unable to plan for preservation action.] **TRAC Reference B1.2** Repository clearly specifies the information that needs to be associated with digital material at the time of its deposit (i.e. SIP)	Metadata for long-term preservation, for example PREMIS.	
	IIPC Web Archiving Metadata Set V2	
	Resource discovery metadata	10/07 Teleconference agreed that this is a long-term accessibility issue.
	Persistent Identifiers	10/07 teleconference agreed this should be added to potential solutions.
	Format identification tools, e.g. DROID	DROID is probably only useful for post-ingest processing for web archives.
	Automatic metadata extraction tools (e.g. NZNL)	
	JHOVE	JHOVE is useful for a limited (but growing) range of formats.
	Date stamping	Date stamping especially useful for repeat crawls and multiple instances of the same resource.
		General Comment: Identifying file formats is likely to increase processing time. Note: Mime type provides an overall indication of what's contained in PANDORA but is insufficient for preservation planning purposes.

Threat/consequence	Potential Standards/Tools/Approaches	Other Comments
7. Idiosyncratic file formation [Idiosyncratic response to preservation processes.] **TRAC references:** **B1.4.1** Repository employs documented preservation strategies. **B3** Preservation planning **B2.7** Repository demonstrates that it has access to necessary tools and resources to establish authoritative semantic or technical context of the digital objects it contains (i.e.access to appropriate international Representation Information and format registries).	QA	QA needs to be largely automated to cope with size of web archive.
	Can submit code to W3C site for validation.	Validating file formats is labour intensive.
	Documentation (especially if choosing to retain idiosyncratic file formation).	Documentation is labour intensive.
		In PANDORA, bad coding needs to be fixed before it will display. This is potentially a bigger issue for Whole Domain Harvests because of the large scale. Thorstein Hallgrimsson noted that there was particular overlap between threats 7-11 and the IIPC Access Working Group [email exchange 21/08/07]
8. Access Chain breaks [unable to render onsite]	Preservation metadata, e.g. PREMIS, IIPC Web Archiving Metadata Set V2.	July teleconference proposed undertaking small-scale testing of selected 'problem' formats to test issues in rendering them. See Also Threat 10
	Global Digital Format Registry (GDFR)	Still under development.
	PRONOM	The Future plans for PRONOM indicate a preservation planning service which holds much promise of filling a current gap for format tools to include risk metrics to enable faster decision-making about priorities for action.
	Automated Obsolescence Notification System **(**AONS)	
	Commercial pdf validators	
	Library of Congress Digital Formats site	
	Migration - at ingest (e.g. Xena) - at risk trigger - on demand (e.g LOCKSS)	Migration may disrupt context, e.g. links may be lost. The 'essence" of a record/document may be lost. The presentation is likely to be changed. Normalisation to XML may increase the size of the file – may become too big for web archives?

8. continued		Experience to date for migration on demand has been primarily for images.
	Emulation/UVC	The KB is currently involved in a project to test emulation as a digital preservation strategy. UVC has been mainly tested on images to date.
	Viewers	
	Durable encoding	e.g. Gladney (2004)
	Software archive	The Cedars project recommended retaining software or at least documentation to enable meaningful access over time.
9. Access chain breaks [unable to render remotely]	Plug-ins for users?	
	Maintaining archive of browsers	
	Design delivery system to enable migration on demand	[Gap: Delivery system to enable migration on demand is not currently available but would be much more effective than maintaining an archive of browsers].
10. Lack of technical experience for preservation action [makes planning difficult]	Testbeds, pilots, focussed trials, recording and sharing information	Conducting tests within the institution help to predict how formats are likely to behave when they are migrated in future. A Study commissioned by the Smithsonian Institute, 2001 tested migration from html to xhtml. Rosenthal et al, 2005b, describe how the LOCKSS system has designed and tested an initial implementation of format migration for Web content that is transparent to readers, building on the content negotiation capabilities of HTTP.
	Measure data integrity [**TRAC Reference A3.8]** See Also Threat 3 Data Corruption + **TRAC Reference B4.4**	A valuable task would be address the question “How do web archives adequately document data integrity?”
	Retaining the original bit-stream.	Reduces risk that migration will lose important information. Both teleconferences agreed that this is an essential contingency plan.

Threat/consequence	Potential Standards/Tools/Approaches	Other Comments
11. Legal Issues [Lack of copyright permission – we can't take copies for preservation purposes without technically breaking the law. Data protection/privacy issues.] **TRAC References** **A3 Procedural, accountability & policy framework** **A5 Contracts, licenses, & liabilities**	Rights policies.	Rights owners may override rights policies unless they are supported by appropriate legislation.
	Legal Risk management plan.	
	Negotiated permissions (e.g. PANDORA).	Labour intensive and impractical for large-scale crawls and harvests.
	Mechanisms to respond to complaints, e.g. offering to remove disputed material.	Runs risk of losing valuable material from the archive.
	Legislation supporting digital preservation, e,g Copyright, Legal Deposit.	Legislation varies from country to country. Achieving and implementing legislation can often be lengthy and may not adequately address digital materials, in particular web content. Usually requires sustained advocacy. Dependent on political issues which may be outside the control or influence of archiving organisations.
12. Inadequate resources a) Organizational Structure and Staffing **TRAC Reference A2**	• Define skills and tasks required • Active professional development program	A 2.1 (staff with adequate skills…) and A2.3 (active professional development program) were two of 8 TRAC Division A criteria selected by NLA as being highest priority for building and maintaining web archives. See Analysis of TRAC on NLA wiki fir further details.
12. Inadequate resources b) Financial Sustainability **TRAC Reference A4**	• Develop short and long-term business processes • Monitor and bridge gaps in funding	Need more information on costing of web archiving activities. The LIFE project is of interest in this respect.
12. Inadequate resources c) System Infrastructure **TRAC Reference C1**		

Attachment 1

Tasks and Responsibilities

Task	Prime responsibility	Progress Report
1. Analyse OAIS and its relevance for web archives	George Barnum	Draft report added to wiki end of October and comments requested.
2. Analyse TRAC for its relevance to web archives	NLA	Document prepared and added to wiki on 18 Sept. The analysis was restricted to Division A: Organisational Infrastructure and selected 8 criteria (subsequently revised to 9 criteria) believed to be highest priority with further work proposed. These include defining skills required, providing feedback on training, models for legal permissions, and costs of web archiving.
3. Provide reports of operational issues and use of tools.	All	KB have prepared a report with some preliminary results using Jhove and Droid. DNB have prepared a report on their experience of using Jhove for the Kopal project. Both documents have been posted to the NLA wiki for further discussion by PWG members. BnF provided a summary report of their project on capturing French election websites. This was added to the NLA wiki on 22 November.
4. Investigate current best practice in disaster prevention and recovery.	Bit Preservation sub-group [no specific responsibility assigned at this stage]	
5. Investigate current best practice in data management	Ditto above	LoC provided a report on Bit-Preservation specifications which was added to the wiki on 21 November.
6. Define how web archives document integrity	Ditto above	
7. Undertake small-scale tests on 'problem' formats to test issues in rendering them.	KB/DNB reported they were undertaking emulation tests	The KB's recent report indicates they are using emulation as their main focus for preserving websites. The Dioscuri emulator V2.0 was released in September 07. DNB added a report from Kopal on their use of Jhove and Droid