

IIPC Preservation Working Group

Report to IIPC Steering Committee

Table of Contents

- I. Executive Summary
- II. Background
- III. Methodology
- IV. Preservation Risks
- V. Standards
- VI. Tools
- VII. Projects
- VIII. Identified Gaps and Recommendations
- IX. Conclusion

- Appendix 1 Terms of Reference for the PWG
- Appendix 2 Table of Threats
- Appendix 3 References for Tools and Standards, Projects, & Further Reading
- Appendix 4 Analysis of TRAC Section A for web archives
- Appendix 5 Analysis of OAIS for web archives
- Appendix 6 KB Reports on use of Jhove and Droid
- Appendix 7 Kopal report on working with Jhove
- Appendix 8 Library of Congress Bit-Level Specifications
- Appendix 9 METS Profile for Web Site Captures
- Appendix 10 BNF summary report. Conducting an event-based web archiving project: the example of the French national elections crawl.

I Executive Summary

At its January 2007 meeting in San Francisco, the Steering Committee agreed to the establishment of a Working Group on Preservation. The IIPC Preservation Working Group held an inaugural meeting in conjunction with an IIPC Steering Committee meeting in April 2007 but conducted most of its work between June-December 2007. Regrettably, it was not possible to schedule any face-to-face meetings so the group relied on a combination of teleconferences and a wiki established by the National Library of Australia to conduct its work.

The PWG focussed primarily on practical issues associated with characterisation of Web archives. It concluded that many existing standards and approaches are applicable to Web archives but in some cases (in particular tools) require some adaptation to work efficiently and effectively to meet the preservation requirements of large-scale Web archives. In terms of standards, OAIS offers a high-level framework, which can be applied to Web archives. The Trustworthy Repositories Audit & Certification: criteria and checklist (TRAC) provides a more detailed implementation of OAIS. The PWG conducted an in-depth analysis of TRAC Section A (Organizational Infrastructure), and identified nine of the twenty four criteria which it believes to be of immediate practical utility for Web archives, and a further four which it believes will become critical in the near future.

PWG members are building up practical experience and have been willing to share the results of this with the group. Some have reports of their experience, which are available as appendices to this report and provide an additional dimension to the work of the group.

Development of Jhove 2 and Droid 2 are seen as welcome indications of a widespread recognition of the need to develop greater efficiency in processing digital objects, a factor of particular relevance for large-scale web archives. Relevant related work was also noted in three projects (Planets, LIFE, and Web at Risk) which are still in progress but which offer scope for useful outcomes as well as fruitful collaboration with the IIPC. The dearth of relevant training widely available was noted as a gap which would benefit from further attention as there is likely to be an ongoing requirement for such training from a wide range of geographically distributed organisations. Opportunities for a more creative interpretation of training, such as travel and staff exchange are also recommended. In terms of preservation requirements, issues associated with the ability to maintain accessibility still require attention and should be regarded as a gap at this stage, though further research may yield more information.

II Background

At its January 2007 meeting in San Francisco, the Steering Committee agreed to the establishment of a Working Group on Preservation. An inaugural meeting of the group was held in conjunction with the Steering Committee Paris meeting in April 2007 and agreed on a statement of work.

The following IIPC members are participants in the Preservation Working Group:

Chair – National Library of Australia [NLA] (represented by Colin Webb, Monica Berko)

German National Library [DNB] (Tobias Steinke)

Library and Archives Canada [LAC] (Pam Armstrong, Steve Sekerak)

Library of Congress [LC] (Martha Anderson, Abbie Grotke, Gina Jones)

National Library of New Zealand [NLNZ] (Steve Knight)

National Library of Norway [NB.NO] (Lars Gaustad)

National Library of Sweden [KB-S] (Eva Muller)

National Library of the Czech Republic [NLCR] (Libor Coufal)

National Library of the Netherlands [KB-N] (Hilde van Wijngaarden)

National Library of France [BnF] (Gildas Illien, Clement Oury)

The National Archives, UK [TNA] (Adrian Brown)

US Government Printing Office [USGPO] (George Barnum)

In addition, Maggie Jones has been employed by NLA for two days a week from June – December 2007, to provide organisational and research support, funded partly by NLA and subject to an application for partial IIPC funding. Gerard Clifton has also been involved in teleconferences and internal NLA workshops.

The IIPC Preservation Working Group was established to investigate and recommend means that web archives may use to achieve their preservation goals. The initial brief for the PWG was to report on currently available standards, practices and approaches that are applicable to the preservation of web archives.

Six work clusters were identified:

Work cluster #1: Characterise the risks to preservation of web archives.

Work cluster #2: Identify available preservation approaches, agreed and proposed practices, and standards which are currently being applied to other kinds of digital collections.

Work cluster #3: Determine the relevance and applicability of these approaches to the preservation of large scale web archives.

Work cluster #4: Identify areas needing research and development or further standardisation work to enable large scale web archives to adopt effective and efficient preservation measures in managing their collections and workflows.

Work cluster #5: Suggest decision-making tools that would help managers of web archives in choosing approaches that are appropriate to their needs and circumstances.

Work cluster #6: Provide a report to the IIPC Technical Committee.

The Terms of Reference for the PWG are at **Appendix 1**.

III Methodology

The initial priority was on work clusters 1-3, characterising the risks to long-term preservation of web archives; identifying preservation approaches, practices, and standards currently being applied to digital collections generally; determine the relevance of the latter approaches for the long-term preservation of large-scale web archives. Work on these focussed on:

- Desk top research.
- Internal NLA workshops to brainstorm threats and potential approaches and standards.
- Preparation of a document (Table of Threats) to use as the basis for further discussion within the PWG and IIPC. This document can be found at **Appendix 2** of this report.

A wiki was established at the NLA to facilitate communication and collaborative working between PWG members.

Work Cluster #5 (Suggest decision-making tools that would help managers of web archives in choosing approaches that are appropriate to their needs and circumstances.) was not addressed in any depth owing to time constraints. However, both the **Table of Treats** and the **TRAC Analysis** provide some guidance.

Seven teleconferences were arranged between July and November. The first six were split into two calls to accommodate different time zones. The final teleconference in November brought together colleagues from Australia, Canada, Europe, and US. It was unfortunately impossible to arrange any face-to-face meetings during the allotted timeframe.

The **Table of Threats** formed the main working tool and further tasks arose out of this document and teleconference discussion. These included:

- A list of references of Tools and Standards, Projects, and Further Reading. These can be found as **Appendix 3** of this report.
- Preparation of an in-depth analysis of Trustworthy Repositories Audit and Certification (TRAC), Part A: Organisational Infrastructure, for its relevance to web archives. This document is available as **Appendix 4** of this report;
- Analysis of OAIS and its applicability to web archives. This document is available as **Appendix 5** of this report.

- Sharing of relevant work undertaken by PWG members. These include internal reports provided by KB on their experience of using Jhove and Droid (**Appendix 6**); the German National Library (DNB) on Kopal's experience of using Jhove (**Appendix 7**); reports provided by the Library of Congress on Bit-Level specification and METS profile for web capture (**Appendix 8 and 9**, respectively); and a summary report provided by the National Library of France (BnF) on their project to collect French national election websites (**Appendix 10**).

Time did not permit these documents to be mined as fully as we would have wished but they nevertheless provide a useful practical backdrop for the work of the PWG and will also supply a valuable resource to facilitate subsequent research.

IV Preservation Risks to be managed

The PWG identified twelve threats which they believed were of most immediate relevance to the long-term preservation of web archives. An initial workshop held at the National Library of Australia proposed potential standards, tools or approaches which might alleviate these threats. The relevant TRAC reference was cited for each threat. The **Table of Threats** was used by the PWG as their primary working document and the full table is available at **Appendix 2**. While there is a degree of overlap in many cases, it was felt more convenient in terms of practical action to further subdivide the threats into three broad groupings:

- Those primarily related to Technical Infrastructure [focussing on Bit-Preservation at this stage]
- Those primarily related to Maintaining Accessibility
- Those primarily related to Organisational Infrastructure

PWG members formed three sub-groups focussing on one of these three aspects.

Threat 1 'Not Taking Action' [Relevant to all three sub-groups]

While in one sense it seems obvious, the PWG thought it worth articulating this broad threat to web archives, which might otherwise be considered 'too hard' for organisations to address. The overarching standards and approaches which seem most suited to mitigating this threat are OAIS and TRAC.

Further work was undertaken by the PWG in preparing a detailed analysis of both of these from the perspective of web archives. See **Appendix 4** (Analysis of TRAC, Section A) and **Appendix 5** (Analysis of OAIS).

The first of the nine criteria selected in the TRAC Analysis of being of particular relevance to web archives was **A1.1** 'Repository has a mission statement that reflects a commitment to the long-term retention of, management of, and access to digital information.' It was however noted that for web archiving, it should ideally be included in a suite of policies, such as collection development, cataloguing, etc. to ensure that it is incorporated into the overall strategic priorities of the organisation.

It was also noted that one of the OAIS mandatory responsibilities 'Follow documented policies and procedures for preservation and access' also relates closely to TRAC criterion **A3.2** 'Repository has procedures and policies in place, and mechanisms for their review, update, and development as the repository grows and as technology and community practice evolve.' This is a relatively early stage of development of web archives so it would be valuable for IIPC members to share any documentation they currently use, to provide a model for others.

Recommendation 1

Collect and collate IIPC Member web archiving preservation policies and procedures to develop model policies and procedures from which all could benefit.

Threats primarily related to Technical Infrastructure [Bit-Preservation]

Table of Threats references: Threat 2 'Viruses'; Threat 3 'Data Corruption'; Threat 4 'Media Failure'; Threat 5 'Disaster'; Threat 10 'Lack of technical experience'

While these threats are all relevant to all digital archives and standard IT good practice can alleviate them, the issues of scale, diversity of content and relative lack of control over content captured as part of large-scale harvests make these particularly challenging for Web archives. It is also worth noting that a joint US/UK Workshop on Digital Preservation proposed that a higher priority be given to bit preservation and in particular, greater use of a geographically distributed bit management and storage infrastructure capable of ensuring bit-level survival.¹

An interesting observation made by the NLA Web Archiving team was the need to run a virus check every time web content is accessed but there can still never be 100% guarantee of web content being virus free so there would always need to be an element of risk management. It was also suggested that it could be useful to retain a virus, once found, and document it. The rationale for this was that viruses are only harmful if they are executed and this could be a relatively safe means of building up documentation about viruses for future reference.

Recommendation 2

Conduct a small study on existing Web archives to determine the true risk from viruses.

Of the nine TRAC Section A criteria considered to be of most relevance to Web archives, **A3.8** 'Repository commits to defining, collecting, tracking, and providing, on demand, its information integrity measurements' is particularly appropriate for this sub-section. The TRAC analysis proposed that 'A task identified for the Bit Preservation and/or Maintaining Accessibility sub group may be to define how web archives adequately document data integrity. This may include defining an

¹ Day, M & Hockx-Yu, H (2006). 'Joint US-UK Digital Preservation Workshop, Washington, D.C, May 7-9 2006.' *International Journal of Digital Curation*, Issue 1, Volume 1, Autumn 2006.p.72 <http://www.ijdc.net/ijdc/article/view/10/9>

acceptable level of loss.’ However this task is also quite challenging and there was insufficient time and resources available to undertake it.

Recommendation 3

Define a methodology for adequately documenting data integrity for harvested websites.

Recommendation 4

Conduct a survey of IIPC members to identify how stewards physically store, maintain and conduct file inventories and fixity audit Web archives.

Recommendation 5

Undertake analysis of existing web archives to determine if it is possible to define the implication of data integrity loss for Web archives, given the nature of this kind of archive. Is it possible to identify the “importance’ or “value” of an object based on repetition, rarity or what function an object serves within the archive?

Threats related to Maintaining Accessibility

Table of Threats references: Threat 6 ‘Inadequate documentation’; Threat 7 ‘Idiosyncratic file formation’; Threat 8 ‘Access chain breaks’ [unable to render onsite]; Threat 9 ‘Access chain breaks’ [unable to render remotely]

The PWG sub-group investigating these aspects raised the following practical implementation issues:

- Tools developed to capture web content more efficiently, such as ARC and WARC, can present challenges in running tools (if they need to be unpacked and then repacked so that tools such as Jhove and Droid can identify and validate formats this causes problems in seamlessly integrating such tools into the workflow);
- Further workflow integration challenges include the need to run different tools, because of the wide variety of formats in web archives (see section on Planets which describes some of the work they are doing in this area);
- Using METS in conjunction with PREMIS provides a rich combination but can also makes file sizes particularly large (there are reports of the METS file being larger than the website!);
- Further work is needed to establish how much time it takes to run tools (**Appendix 6** includes a test undertaken by the KB to compare running times for Jhove and Droid);

These comments also reflect a broader issue, of the relatively early stage of tools development which currently makes them difficult to integrate seamlessly into workflows. It is however a problem which has already been noted and developments such as Droid 2 and Jhove 2 are seeking to address the simultaneous need for greater sophistication in tools as well as greater efficiencies. See also sections **VI** (Tools) and **VII** (Projects) for further discussion of these developments. At a very practical level, it would be useful to conduct a

small study which would help to achieve a better understanding of the benefits and barriers of two well known standards and approaches, PREMIS and METS.

Recommendation 6

Conduct a small study on the pros and cons of using PREMIS and METS for web archives.

The Planets project proposes “innovative solutions for Preservation Action tools which will transform and emulate obsolete digital assets”. Arguably, the uniqueness of Web archives is that viewing tools (browsers, etc.) should render content over time and not require emulation or transformation. There is over ten years worth of archives available to test current viewing tools capabilities to render earlier archives to allow an identification of potential risk due to time and evolution of browsers.

Recommendation 7

Conduct a study on existing Web archives to attempt to identify if there is a risk that future viewing tools will not render objects in Web archives or what kind or characteristics of objects are more at risk.

Threats related to Organisational Infrastructure

Table of Threats references: Threat 11 ‘Legal Issues’; Threat 12 ‘Inadequate resources’

Although there are only two threats identified for this sub-group, they represent such a major challenge for all organisations that much effort of the PWG focussed on these. The analysis of TRAC, Part A provided a useful platform for expanding on potential solutions to key organisational challenges.

Legal Issues

Of the nine TRAC criteria selected as being of highest immediate relevance and priority for web archives, two, **A3.3** (‘Repository maintains written policies that specify the nature of any legal permissions required to preserve digital content over time, and repository can demonstrate that these permissions have been acquired when needed.’) and **A5.5** (‘If repository ingests digital content with unclear ownership/rights, policies are in place to address liability and challenges to those rights’) are of direct relevance to the legal issues facing web archives.

The analysis identified barriers to achieving **A3.3** as ones of scale, the diversity of legal jurisdictions, and the varying levels of risk different organisations may be willing to take. Follow-up work proposed by the PWG includes:

Defining what legal permissions are required and providing some models of where these are adequately addressed by existing legislation or other mechanisms. One example of this is the welcome trend in some legal deposit legislation is to include statements relating to the need for the libraries to gain unimpeded access to material they collect. Clauses relating to this issue are included in Canadian, Danish, French, German, Icelandic, and New Zealand legislation. Further research could provide

other examples of specific preservation needs and models of how these have been addressed.

Recommendation 8

Collect and collate relevant clauses in legal deposit legislation which support preservation of web archives.

Recommendation 9

Undertake a survey of IIPC members to identify a) what preservation-specific legal challenges they face for web archives; b) what relevant legislation is in place [other than legal deposit legislation] which either supports or impedes this requirement ; and c) what other mechanisms (such as policy statements, risk management plans etc.) they have in place to deal with the challenges.

TRAC criterion **A5.5** was not discussed in depth by the PWG but seems to be a useful general recommendation for Web archives.

Recommendation 10

If repository ingests digital content with unclear ownership/rights, policies are in place to address liability and challenges to those rights' [TRAC **A5.5**]

Inadequate Resources

This is such a major issue for all organisations that it was further subdivided into three sub-sections, 'Organisational structure and staffing'; 'Financial sustainability', and '**System Infrastructure**' (not further addressed or discussed). In the time period for the PWG, most effort was focussed on the first two of these, 'Organisational structure and staffing' and 'Financial Sustainability'. Relevant criteria were identified in the TRAC analysis and some follow-up work was done, as follows:

Organisational Structure and Staffing

A2.1 'Repository has identified and established the duties that it needs to perform and has appointed staff with adequate skills and experience to fulfil these duties.'

Barriers to achieving this criterion included defining specific skill sets required. The PWG Organisational Issues sub-group gathered examples of recent job adverts which are provided as an attachment to the TRAC analysis [**Appendix 4**]. These provide some models of how a range of organisations are defining their requirements for digital archives generally if not specifically for web archives. Defining what skills and experience are required is also helpful for structuring professional development training, as well as for recruitment purposes.

A2.3 'Repository has an active professional development program in place that provides staff with skills and expertise development opportunities.'

The need for continuous professional development seemed to the PWG to be crucial in such a rapidly changing environment. Recent training workshops announced by the European Archive have been noted and it is hoped that similar programmes will

be developed elsewhere to enable geographically dispersed access to structured training. Given the topic, it may also be feasible to develop a web-based training programme. The TRAC analysis also alluded to the need for a broader interpretation of what might constitute training, important for such a relatively new area.

Examples of these include opportunities provided by IIPC events for members to transfer skills and experience through both formal and informal mechanisms, such as staff exchanges. A related recommendation made in the TRAC analysis is for organisations to consider including travel as a legitimate training cost as there are still a relatively small number of organisations actively involved in web archiving and the ability to travel to them to discuss specific issues and to facilitate peer to peer skills transfer could be of great value in helping to transfer knowledge and experience more widely. The dearth of suitable training programmes is noted as a current gap.

Recommendation 11

Undertake further analyses of recent job adverts, including those collated in Attachment 1 of the TRAC Analysis, to help develop a set of core skills, experience and qualifications required for web archives.

Recommendation 12

Web Archives require a range of training mechanisms, including widely accessible training programmes which cover the full range of requirements.

Financial Sustainability

A4.1 'Repository has short- and long-term business planning processes in place to sustain the repository over time.' This is closely related to **A4.5** 'Repository commits to monitoring for and bridging gaps in funding.'

Further work on these criteria concentrated on an analysis of the relevant web archiving case study from the final report of Stage 1 of the LIFE Project² (See Also **Section VII**, Projects). The potential of this project to provide more concrete predictions of web archiving costs is an important first step in building a sustainable business case. The case study selected for the LIFE project was the British Library's contribution to the UK Web Archiving Consortium (UKWAC) and is therefore a selective web archive, similar to the National Library of Australia's PANDORA.

Figure 1 illustrates the findings of the LIFE project, when estimating costs over 20 years for this selective archive. This assumes that preservation will be the highest single category of costs over this timeframe, with metadata and access costs assumed to be so modest, they have been combined for the pie chart. As the researchers noted, several cost elements needed to be estimated as there is little or no empirical evidence for some of them over this period of time. However, it provides a useful baseline from which to conduct further comparative research.

² McLeod, R, Wheatley, P & Ayris, P (2006). Lifecycle information for e-literature; full report from the LIFE project. <http://eprints.ucl.ac.uk/archive/00001854/01/LifeProjMaster.pdf>

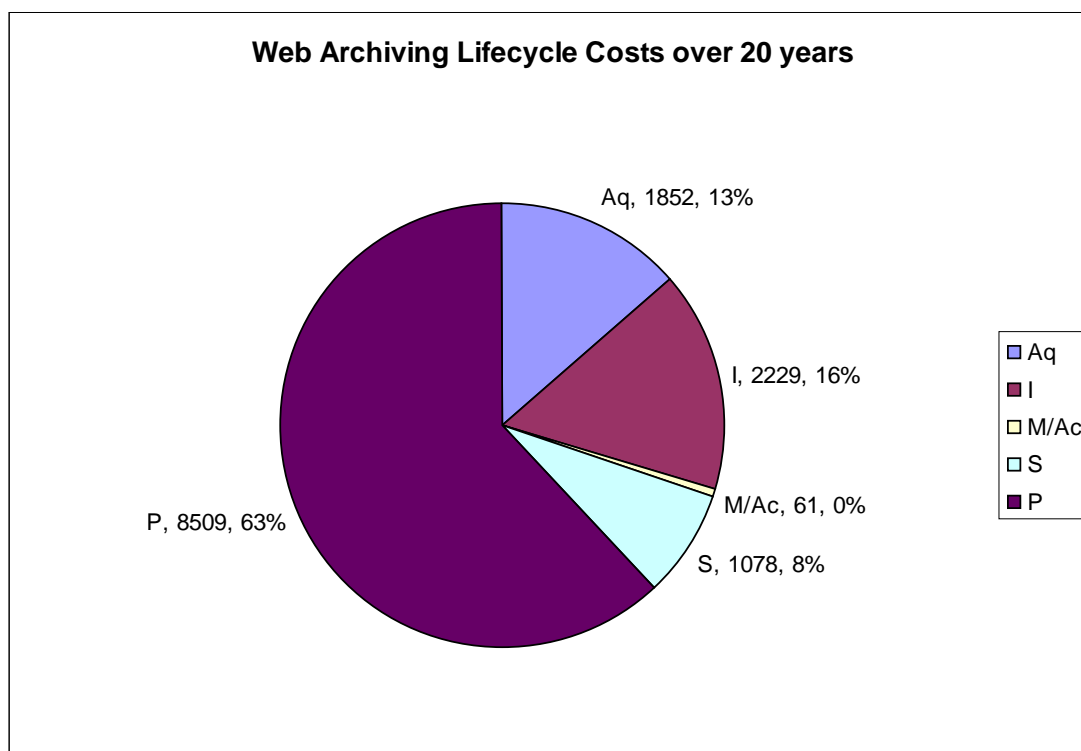


Figure 1

A report conducted by the BnF (**Appendix 10**) also provides helpful indicators of costs for their election web archiving project. The second stage of the LIFE project will be applying the methodology to different case studies and so will not be testing the formula on web archives. Contact has been made with LIFE project staff and it has been agreed that it would be valuable to conduct further tests of the methodology, comparing both other selective archives as well as large-scale domain harvests with the initial results from Stage 1 of the LIFE project³.

Additional resources would be required to undertake this but PWG members would be in a strong position to provide case studies for the project.

Recommendation 13

Web archives should develop short and long-term business planning processes to sustain the web archive over time. [Note: This is adapted from TRAC A4.1]

Recommendation 14

Seek funding [possibly from EU] to conduct further tests on predicting costs for web archives, building on the LIFE Project work.

³ It should also be noted that the LIFE formula has subsequently been slightly modified, taking on board feedback and additional thoughts from the project team. However the overall predictions for web archiving costs made using the original formula should not be affected. The LIFE model V1.1 is available from <http://eprints.ucl.ac.uk/archive/00004831/01/4831.pdf>

V Standards

While on the one hand there are numerous standards and approaches applicable to Web archives, there are a much smaller number which specifically relate to their long-term preservation, the focus of this working group. Of the latter category, the group concentrated primarily on analysing OAIS and TRAC in some depth. The conclusion drawn was that both of these applied to web archives, albeit with some modifications. They both provide a useful benchmark for Web archives, as well as offering a means of identifying significant gaps. During the timeframe of the PWG, emphasis was given to Section A of TRAC, relating to Organizational Issues. It would also be worth analysing Parts B and C in more depth, possibly in relation to some of the technical recommendations made in this report. Other international standards and approaches which appear to apply equally well to web archives are WARC, PREMIS and METS. Lack of time prevented further work on these but both are the subject of a recommendation for further work. In addition, **Appendix 9** provides a METS profile for web site captures.

VI Tools

Not surprisingly, the emphasis on tool development to date has tended to be on tools which support the more efficient and effective capture of relevant web content. Tools supporting long-term preservation tend to be less well developed though two tools for identifying and validating file formats, Jhove, and Droid, are being increasingly used by digital archives generally and in both cases, the latest version promises greater sophistication and utility for web archives. Two PWG members (KB and DDB) provided reports on their experience of using Jhove and Droid and can be found at **Appendix 6** and **7**, respectively.

VII Relevant Projects

In addition to the LIFE project, discussed in **Section IV**, two other projects were identified as being of particular relevance and interest to the PWG. These are Planets and the Web at Risk, described in more detail below:

Planets (**P**reservation and **L**ong-term **A**ccess via **N**etworked **S**ervices), an EU funded integrated project led by the British Library.

Though it is at an early stage (it began in 2006 and is due to be completed in 2010), there is clearly potential for Planets to develop tools which can assist the digital preservation community generally, including web archives. A recent presentation added to the Planets website articulates some of the concerns also expressed by PWG members⁴. These include the difficulties of seamlessly integrating the use of tools in workflows, and the potential need to use different tools for different requirements, for example one tool to identify the file format, another to validate it, another to extract relevant metadata etc.

⁴ Sharpe, R (2007). Automated Characterisation Framework
http://www.planets-project.eu/docs/presentations/Planets_Tools-and-Trends_RobertSharpe.pdf

A PWG member, Hilde Wijngaarden, has acted as a liaison between the Planets project and the PWG as the KB is a Planets partner. This has enabled a richer understanding of the potential benefits of Planets to assist with some of the issues relevant to web archives. The potential for PWG members to act as test beds for Planets tools was seen as being very valuable, both for the project and also for the PWG members, who have specific practical requirements which the tools may help with.

Web At Risk, a project coordinated by the California Digital Library, the first stage of which was supported by the Mellon Foundation and the subsequent stage is NDIIPP funded. Liaison with this project was facilitated by PWG members Martha Anderson and George Barnum. Tracy Seneca was invited to join one of the PWG teleconferences to provide further background and detail. This project was of interest to PWG members because it is developing a central preservation service to support the preservation of web-based government information. Like the Planets project, Web at Risk is not yet completed (it is funded until 2009) so it would be useful to maintain an active watching brief on this project as well.

Recommendation 15

Continue monitoring the LIFE, Planets, and Web at Risk projects and ensure ongoing collaboration between them and the IIPC.

VIII Identified Gaps and Recommendations

The gaps identified by the PWG reflect the emphasis on threats regarded as the highest priority and those which offer some immediate (or at least short-term) practical resolution. They also reflect the initial concentration on organisational issues.

While there are undoubtedly gaps in terms of tools which can automate tasks required for characterisation and preservation planning – in particular the identification and validation of file formats, work proceeding on updating existing tools and through projects such as Planets, offers the potential to significantly improve this situation. Opportunities to influence and actively participate in such developments should be sought by all IIPC members. The next IIPC General assembly, hosted by the National Library of Australia in 2008, will also provide an opportunity for further refinement of practical issues associated with current tools.

Other gaps related to organisational issues, such as the absence of detailed costs and business models, staff recruitment and ongoing professional development. These are the subject of some of the recommendations below.

The recommendations can be divided into general recommendations relevant to all web archives, and those specific to IIPC to undertake further work on.

General recommendations for all Web archives

Recommendation 10

If repository ingests digital content with unclear ownership/rights, policies are in place to address liability and challenges to those rights' [TRAC A5.5]

Recommendation 12

Web Archives require a range of training mechanisms, including widely accessible training programmes which cover the full range of requirements.

Recommendation 13

Web archives should develop short and long-term business planning processes to sustain the web archive over time. [Note: This is adapted from TRAC A4.1]

Recommendations to IIPC for further work:

Recommendation 1

Collect and collate IIPC Member web archiving preservation policies and procedures to develop model policies and procedures from which all could benefit.

Recommendation 2

Conduct a small study on existing Web archives to determine the true risk from viruses.

Recommendation 3

Define a methodology for adequately documenting data integrity for harvested websites.

Recommendation 4

Conduct a survey of IIPC members to identify how stewards physically store, maintain and conduct file inventories and fixity audit Web archives.

Recommendation 5

Undertake analysis of existing web archives to determine if it is possible to define the implication of data integrity loss for Web archives, given the nature of this kind of archive. Is it possible to identify the "importance" or "value" of an object based on repetition, rarity or what function an object serves within the archive?

Recommendation 6

Conduct a small study on the pros and cons of using PREMIS and METS for web archives.

Recommendation 7

Conduct a study on existing Web archives to attempt to identify if there is a risk that future viewing tools will not render objects in Web archives or what kind or characteristics of objects are more at risk.

Recommendation 8

Collect and collate relevant clauses in legal deposit legislation which support preservation of web archives.

Recommendation 9

Undertake a survey of IIPC members to identify a) what preservation-specific legal challenges they face for web archives; b) what relevant legislation is in place [other than legal deposit legislation] which either supports or impedes this requirement ; and c) what other mechanisms (such as policy statements, risk management plans etc.) they have in place to deal with the challenges.

Recommendation 11

Undertake further analyses of recent job adverts, including those collated in Attachment 1 of the TRAC Analysis, to help develop a set of core skills, experience and qualifications required for web archives.

Recommendation 14

Seek funding [possibly from EU] to conduct further tests on predicting costs for web archives, building on the LIFE Project work.

Recommendation 15

Continue monitoring the LIFE, Planets, and Web at Risk projects and ensure ongoing collaboration between them and the IIPC.

IX Conclusion

The number of recommendations for further action is indicative of the additional effort the Preservation Working Group believes is required at this very early stage of the investigation of long-term preservation of Web archives. However, despite noting the need for further work, it is clear that there is a growing corpus of practical experience which has already achieved a degree of consensus around some practical issues, and also provides an excellent foundation for fruitful collaboration. Issues relating to the ability to maintain accessibility into the future will provide particular scope for further research. This report represents the results of the deliberations of the Preservation Working Group which has taken place primarily over the past six months and has relied overwhelmingly on the willingness of PWG members to dedicate time from their busy schedules to contribute to the PWG work programme. The PWG now invites further feedback from the IIPC Technical Committee on how to make further progress in this important area.