Report on Preservation Tools (R 1.3.2)

Version 12/14/2011
Workpackage 1
Responsible Partner SUB

DARIAH-DE
Aufbau von Forschungsinfrastrukturen für die e-Humanities

This research and development project is / was funded by the German Federal Ministry of Education and Research (BMBF), fund number 01UG1110A to M, and managed by the Project Management Agency the German Aerospace Center (Deutsches Zentrum für Luft- und Raumfahrt, PT-DLR).

SPONSORED BY THE

Federal Ministry of Education and Research

Project: DARIAH-DE: Aufbau von Forschungsinfrastrukturen für die e-Humanities
BMBF Fund number: 01UG1110A to M
Duration: March 2011 till February 2014

Document status: final
Dissemination level: public
Authors:

Andreas Aschenbrenner, SUB
Patrick Harms, SUB

Revisions:

Date	Author	Comments
12/14/2011	Aschenbrenner, Harms	First version based on Aschenbrenners draft
01/17/2012	Harms	Final changes and corrections

Table of Contents:

1. Introduction 4

2. The Preservation Context 4

3. Technical Properties of Existing Tools 6

4. Existing Tools 8

5. Existing Frameworks for Preservation Tools 9

6. Conclusions 11

1. Introduction

The preservation community has been creating strategies, concepts and tools for preservation activities for many years[1]. Since then, numerous activities with diverse backgrounds have created preservation systems (policies, organisational workflows, as well as technologies) and actual preservation tools in a highly decentralised manner. Today it is widely acknowledged that preservation is more an organisational than a technical challenge; moreover, there are no actual preservation tools, but only tools and technologies that may support a preservation strategy in a specific context.

This report aims to learn from the experiences in the preservation community in two ways:

(1) Due to the highly decentralised nature of the preservation community, there may be concepts for technical interoperability across diverse tools that are valuable for the similarly decentralised environment in the arts and humanities (AH). However, the analysis of technical interoperability concepts in the preservation community shows that there is no ultimate answer to technical interoperability and that only local islands of interoperability can be fostered through shared conventions.

(2) Several DARIAH user groups have an interest in effective preservation (including archives containing AH research data, research networks, individual researchers) and are looking for a specific toolset for preservation. However, as pointed out above, preservation is not a (purely) technical challenge and there is no definite toolset for preservation, hence this report rather encourages the creation of guides to tailor a preservation approach to a specific context.

2. The Preservation Context

Preservation of digital objects has numerous aspects and perspectives.[2] The preservation community is very diverse; institutions performing preservation actions span a wide area of different business contexts (from national archives, to corporate archives, to research networks and individual researchers with actual obligations to preserve their data). Since we cannot and do not intend to cover all preservation-related issues, this chapter aims to set the scope by outlining some organisational and architectural aspects in preservation.

---

[1] While actual preservation activities have been performed for many decades, one of the first comprehensive reports and analyses on digital preservation is from 1996:
Preserving Digital Information. Report of the Task Force on Archiving of Digital Information. 1996.
http://www.clir.org/pubs/reports/pub63watersgarrett.pdf

[2] Colin Webb. Guidelines for the Preservation of Digital Heritage. UNESCO Report, United Nations Educational, Scientific and Cultural Organization, Paris, March 2003.
http://unesdoc.unesco.org/images/0013/001300/130071e.pdf.

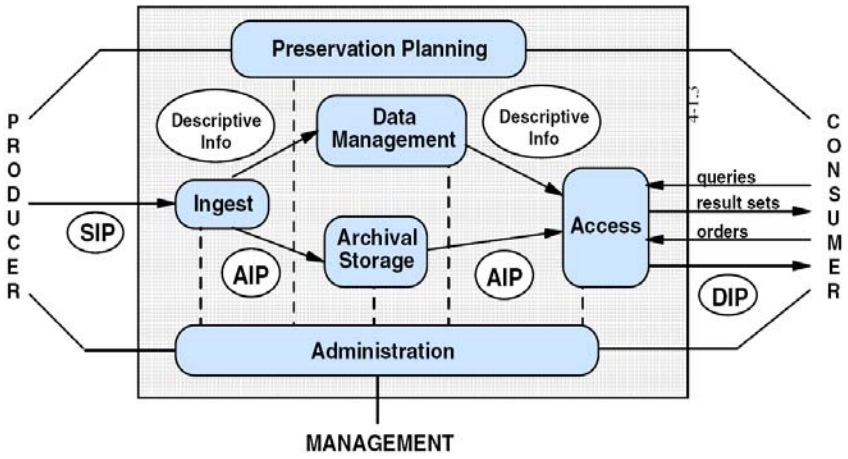

Figure 2.1: OAIS Functional entities

One of the key standards in digital preservation is the OAIS ISO standard[3], which defines a terminology and identifies functional units related to digital preservation. It also makes clear that digital preservation is primarily an organisational challenge (including policy, roles and responsibilities, financial sustainability, and procedural feasibility)[4], and cannot be solved by technology alone. Nevertheless, several functions of a preservation system as defined by the OAIS (cf. Figure 2.1) can be supported through technology. This report presents some existing tools in the OAIS functional entities.

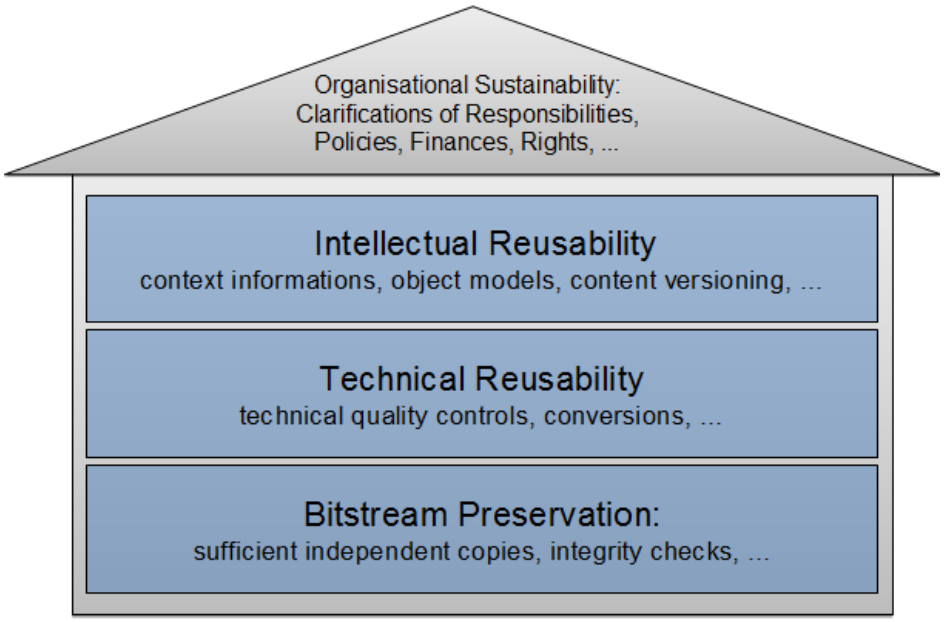

Figure 2.2: Levels of abstraction and related activities for preservation

[3] Reference Model for an Open Archival Information System (OAIS). Recommendation for Space Data System Standards, CCSDS 650.0-B-1. Blue Book. Issue 1. Washington, D.C.: CCSDS, January 2002. [Also published as ISO 14721:2003.]

[4] Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist. http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying

By looking at the digital objects to be preserved rather than the preservation systems, we distinguish different levels of abstraction (cf. Figure 2.2): the bitstream, technical reusability, and intellectual reusability levels. Each of these levels involves different tasks and may involve different agents to conduct and oversee these tasks.[5]

This report focuses primarily on tools on a technical reusability level. The bitstream level is close to hardware-related issues and is therefore often handled by data centres through standard operating procedures.[6] On the other hand, intellectual reusability is strongly influenced by the particular business resp. research context, and hence can hardly be transferred e.g. from one research project to another.[7] Technical reusability, however, involves several tasks that are similar across disciplines and can (to some degree) be supported by generic tools[8].

This report aims to

(1) identify possible technical activities to support preservation activities in arts and humanities (AH) research, and to

(2) learn from the preservation community, how tool interoperability across a diverse and distributed community can be achieved.

3. Technical Properties of Existing Tools

This chapter aims to categorise existing tools. Such a categorisation may then guide the survey of existing tools in the following chapter, as well as the analysis of interoperability in preservation tools after that. However, rather than succeeding in defining a simple categorisation, this chapter will underline the breadth in goals and diversity in existing implementations of preservation tools through several distinct categories for preservation tools.

From an organisational perspective, preservation tools may support one (or many) of the OAIS functional entities mentioned above (cf. Figure 2.1): Ingest, Preservation Planning, Data Management, Archival Storage, Administration, and Access.

In each of those areas, tools vary with regard to their dependencies and as to how they are embedded in the respective system context. For example, the preservation policy of a large repository for historic texts may favour formats like TEI/XML and PDF/A, over Microsoft Word and other proprietary formats. It therefore converts incoming objects on ingest into those standard formats, using fairly generic tools for format conversion. One step further in the ingest workflow, the repository validates the converted file with the user, creates the SIP (Submission Information Package, cf.

---

[5] Andreas Aschenbrenner, Harry Enke, Thomas Fischer, Jens Ludwig: Diversity and Interoperability of Repositories in a Grid Curation Environment. In: Journal of Digital Information, Vol 12, No 2 (2011). http://journals.tdl.org/jodi/article/view/1896

[6] WissGrid: Bitstream Preservation - Bewertungskriterien für Speicherdienste. March 2011. http://www.wissgrid.de/workgroups/ap3/2011-03-08--bitstream-preservation.pdf

[7] Requirements may differ even within the same discipline or within a single institution. Metadata, data formats, retention schedules, etc. essentially depend on a specific research question as well as on the procedures agreed in a specific research activity.

[8] Although these tools may need to be embedded into the specific organisational and technical context, and may need to be configured to satisfy the requirements of the preservation policy in place.

OAIS) and registers the SIP with the repository. Those latter steps depend on the mode of interaction with the user, the requirements for metadata - amongst other - when building the SIP, and the protocols available for transfer to the repository. In other words, all these steps need to be suited (individually) to their respective organisational and technical context, before chaining them together to establish the complete ingest workflow suited for the repository for historical texts.

Referring to their technical architecture, a preservation tool may be algorithmic or be based on a reference database. For example, a tool for extracting technical metadata from PDF files may be entirely algorithmic and can be embedded in numerous technical contexts. On the other hand, a format registry that contains file format specifications for future reference is based on a central database. Due to the effort of sustaining such a central database, only few format registries are expected to be established and used for global reference.[9] Added-value services of such reference databases may - although algorithmic - be tied to the reference database.

With regard to user interaction, preservation tools may be interactive, semi-automatic or automatic. For example, tools for planning are predominantly interactive tools, whereas a tool for format conversion can be widely or entirely automatic.

Please note, that there are only few tools that are entirely automatic. For example, when bulk converting millions of files from a legacy format to a current format, there are bound to be issues with some of the files. Even if the tool does not crash due to a (minor) error in the file or a bug in the tool, which may need to be treated individually (depending on the quality requirements of the archive), random checks on the results may identify unexpected behaviour. Similarly, there are hardly tools that cover a specific functionality comprehensively (i.e. THE tool for X in all preservation contexts). For example, even if only converting from one specific format to another, conversion tools may vary in their assumptions, with regard to the significant properties they cover for those data types, and how they integrate with different system contexts.

Drilling further into the category of automatic preservation tools, we may distinguish between installations that are local in the archive[10], and those where data are streamed to a remote site. Even commercial service providers are conceivable for streaming services.[11] A decision between "data to the service" or "service to the data" depends on aspects like the size of the data and expected transfer latency, as well as the technical architecture of the archive.[12]

Last but not least, and apart from the functional aspects mentioned above, existing preservation tools may stem from various backgrounds with different promises re-

---

[9] Stephen L. Abrams and David Seaman, “Towards a Global Digital Format Registry,” World Library and Information Congress: 69th IFLA General Conference and Council, Berlin, August 1-9, 2003 <http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf>.

[10] A service may of course be "local" along various dimensions. In this case we refer to a strong definition of "local" where a tool is (1) part of the organisational context of the archive, (2) part of its technical architecture and (3) physically close to the archive to avoid transfer latency.

[11] Organisations like the Open Planets Foundation (www.openplanetsfoundation.org) may eventually offer such cloud-based streaming services.

[12] For example, for matters of security and trust, archives may choose to prevent services to be executed within the archive. Depending on the technical architecture and the available resources in the datacentre(s) hosting the archive, a service may then be forced to be placed "close to" the archive or to a remote site.

garding their sustainability. Moreover, some functions in typical preservation systems may be required in numerous contexts, and hence suitable tools may or may not be written with the goal of preservation. For example, tools for format conversion have existed before they were required for preservation actions, and the most suitable tool for a particular function may continue to be created without preservation in mind.

4. Existing Tools

The previous chapter identified several dimensions along which preservation tools can be categorised. This chapter surveys some actual tools along those dimensions.

Some very well known tools used in preservation environments include the following. This selection of tools is purely exemplary to illustrate the categorisation of tools a-long the dimension.

- JHove[13], format validation
  - *OAIS function*: Ingest
  - *technical architecture*: automatic, streaming tool
  - *organisational background*: sustained by the preservation community through successive projects, due to its popularity
- Pronom[14], format registry
  - *OAIS function*: Preservation Planning (technology watch)
  - *technical architecture*: reference database with some emerging added-value services
  - *organisational background*: sustained by UK national archives
- Plato[15], preservation planning tool
  - *OAIS function*: Preservation Planning
  - *technical architecture*: online, interactive, expert system
  - *organisational background*: academic effort
- numerous tools for format conversion
  - *OAIS function*: Ingest, Access
  - *technical architecture*: command line or streaming
  - *organisational background*: commercial and non-commercial, e.g. part of regular Linux distributions

[13] JHove, JSTOR/Harvard Object Validation Environment. http://hul.harvard.edu/jhove/

[14] Pronom. UK National Archives. http://www.nationalarchives.gov.uk/PRONOM/

[15] Plato. http://www.ifs.tuwien.ac.at/dp/plato/intro.html

While this only gives a brief idea of existing preservation tools, it re-emphasises the breadth of existing services with regard to their functions, technical architecture, and organisational backgrounds.

Identification and evaluation of preservation tools is an ongoing task, since relevant tools may emerge and be discontinued at all times. Therefore, rather than attempting to produce a complete evaluation ourselves, this index refers to some recent (and ongoing) surveys:

(1) by the EU-project SCAPE (2011)[16]

(2) by the NDIIPP infrastructure of the US Library of Congress[17]

(3) by the JISC project CAIRO (2007)[18]

(4) by the JISC project AQuA (in cooperation with the Open Planets Foundation)[19]

5. Existing Frameworks for Preservation Tools

Several attempts were previously made to connect existing preservation tools, and even to provide a framework that covers all (or rather: many) technical support tasks in functional entities of the OAIS. This chapter aims to identify some of them, evaluate whether and why they are successful, and looks for potential lessons to be learnt from them.

Today, it is widely accepted that the OAIS is not a specification for implementing a digital archive, but rather a checklist of functions that may or may not be relevant in a specific organisational context. Similarly, interoperability frameworks in digital preservation need to be

(1) adaptable to the specific organisational environment and which functions are actually needed, and

(2) flexible as to which particular tool implementations are used[20].

---

[16] SCAPE (SCAlable Preservation Environments). e.g.

- Identification and selection of large-scale migration tools and services. June 2011. http://www.scape-project.eu/wp-content/uploads/2011/09/SCAPE_D10.1_KEEPS_V1.0.pdf
- Evaluation of characterisation. Part 1: Identification. End 2011. http://www.openplanetsfoundation.org/system/files/SCAPE_PC_WP1_identification21092011_0.pdf

[17] NDIIPP Partner Tools and Services Inventory. National Digital Information Infrastructure and Preservation Program (NDIIPP). http://www.digitalpreservation.gov/partners/resources/tools/index.html

[18] Cairo tools survey: a survey of tools applicable to the preparation of digital archives for ingest into a preservation repository. 21 May 2007. http://cairo.paradigm.ac.uk/projectdocs/cairo_tools_listing_pv1.pdf

[19] AQuA Mashup Tool List. http://wiki.opf-labs.org/display/AQuA/AQuA+Mashup+Tool+List

[20] E.g. for each action X, there may be several suitable tools. An interoperability framework does not define the "right" tool of all those available, but enables the use of any of those.

Approach 1, Object-Centric

Central to the approach by the TIPR project (Towards Interoperable Preservation Repositories)[21] is an object format called the Repository eXchange Package (RXP). Rather than focusing on services, APIs or protocols, TIPR focuses on the data. Its assumption is that whatever technical properties the tools may have, they essentially have to understand the data for processing and also should produce data that is understandable for other tools.

Understanding the data is of course an important element of interoperability, and may cover much of the communication between preservation repositories (i.e. between distinct OAIS's). However, it does not explain how/when data are passed through chains of tools and many of the functions within a single OAIS.

Approach 2, Technology Standards

One widely cited approach is an Interoperability Framework developed by the EU-Project PLANETS[22]. It comprehensively covers aspects of the technical architecture, defines common APIs, a digital object model, as well as a workflow engine, and is agnostic to storage. Its specifications and implementations (e.g. service registry) are designed to be flexible and evolve over time.

The technology framework still remains to attract supporters and implementers, to ensure a living, sustainable ecosystem. A similar, web-service based approach has been suggested in 2005,[23] but has been discontinued in the meantime.

Approach 3, Simple Services

Like data need to be migrated from a legacy format to a current one in recurring migration cycles, also systems need to be migrated to new technology platforms from time to time. Therefore, any actual implementations and also any interoperability standards may become outdated at some point. Another approach for interoperability of preservation tools anticipates system migrations by emphasising simplicity. The UC3 curation services[24] assume that decomposing preservation tools into atomic actions, specifying particularly simple APIs (based on e.g. POSIX and HTTP/REST), and developing bare minimum implementations will bring them a long way towards a stable preservation environment and "migratability".

Overall, there is no silver bullet to interoperability, and none of the existing interoperability frameworks have drawn enough community to serve as a reference. There

---

[21] Priscilla Caplan, William Kehoe, Joseph Pawletko: Towards Interoperable Preservation Repositories: TIPR. In: International Journal of Digital Curation. Vol 5, No 1 (2010). http://www.ijdc.net/index.php/ijdc/article/view/145

[22] Planets: Consolidated Release and Documentation. May 2010. http://www.planets-project.eu/docs/reports/Planets_IF-D11_ConsolidatedReleaseDocumentation.pdf

[23] Jane Hunter, Sharmin Choudhury: PANIC - An Integrated Approach to the Preservation of Composite Digital Objects using Semantic Web Services. Presented at the 5th International Web Archiving Workshop (IWAW05), Vienna 2005. http://iwaw.europarchive.org/05/papers/iwaw05-hunter.pdf

[24] University of California Curation Center. Curation Services. http://www.cdlib.org/services/uc3/curation/

are only a couple of properties that can be conducive to interoperability in large, decentralised environments:

- shared concepts of functions and data models are first
- simplicity with regard to technology: to make it easy to get in and easy to get out
- documentation: to make it easy to get in and find out
- design for change: data models and tools will change over time, and you may want to support and record the change
- accept that you cannot dictate any standard or technology, yet be bold to choose and use one - there will be a system migration coming up anyway
- eventually, interoperability is where people make the effort to connect: go with standards and community

6. Conclusions

This report started with two guiding questions, to

(1) identify possible technical activities to support preservation activities in arts and humanities (AH) research, and to

(2) learn from the preservation community, how interoperability of preservation tools across a diverse and distributed community can be achieved.

With regard to the latter (2), the previous chapter recognised that there is no silver bullet to interoperability of tools in a large and diverse community, and distilled a number of recommendations from existing interoperability frameworks. By accepting that interoperability is always local and temporary (i.e. also interoperability is subject to change), the task for achieving interoperability transforms from a technical task to an organisational one: build for change in the technology framework, and build a large user community. A pioneering innovation project may need to go where no-one has gone before, but infrastructure foremost needs to be pragmatic and move along with the masses.

Also with regard to (1) possible technical activities to support preservation activities in arts and humanities (AH) research, this report has no simple answer. On the positive side: the concepts and tools created by the preservation community apply to AH as well, so DARIAH does not need to support any particular tools to enable preservation in the AH in the first place. However, preservation remains an important issue when it comes to its implementation, and DARIAH may have a role in advising its user base in tandem with the preservation community. Some of the questions DARIAH may need to address include:

- (archives containing AH research data)[25]
  Which particular software setup makes sense for my organisation and how

[25] The preservation community has devised guides and checklists like TRAC to help archives build their preservation strategy. Therefore, DARIAH can be more specific to the situation in the AH and help establishing links to AH infrastructure.

can I ensure interoperability with DARIAH (e.g. PID, AAI)? (cf. VCC1 Archive-in-a-Box)

- (research networks)
  How can I create a preservation plan for the gigabytes of data we aim to create over the next couple of years?
- (researchers)
  Where can I find a preservation repository that is suitable for my data and why should I trust it to preserve my data over the next 10 years?

Most of these questions involve tasks in VCC3, or at least require collaboration between VCC1 and VCC3. This applies particularly for the creation of an "Archive-in-a-Box" (VCC1), which can - as this report showed - hardly be a single software solution that covers preservation in the AH comprehensively, but will rather produce a requirement matrix and a set of How-Tos tailored to archetypical scenarios.

In this sense, this report failed to identify simple, generic technical solutions for preservation, because we argue they do not exist. However, hopefully this report succeeded in encouraging AH users in boldly moving on.

Although this report focused on the level of technical reusability (cf. OAIS functional entities, Figure 2.1), DARIAH is currently active on all three levels. On a bit-preservation level, DARIAH-DE is currently building a simple, generic service together with national data centres; once this service is in production it may become a reference for other data centres as well. On an intellectual level, VCC3 is devising data management guides and meta/data standards that may guide AH users in the future.