DCC | Digital Curation Reference Manual

Instalment on
“Preservation Scenarios for Projects Producing Digital Resources in the Area of Digital Humanities”

Milena Dobreva

November 2013
Version 1.0
Legal Notices

The Digital Curation Reference Manual is licensed under a Creative Commons Attribution - Non-Commercial - Share-Alike 2.0 License.

© in the collective work - Digital Curation Centre (which in the context of these notices shall mean one or more of the University of Edinburgh, the University of Glasgow, and the University of Bath and the staff and agents of these parties involved in the work of the Digital Curation Centre), 2005.

© in the individual instalments – the author of the instalment or their employer where relevant (as indicated in catalogue entry below).

The Digital Curation Centre confirms that the owners of copyright in the individual instalments have given permission for their work to be licensed under the Creative Commons license.

Catalogue Entry

<table>
<thead>
<tr>
<th>Title</th>
<th>DCC Digital Curation Reference Manual Instalment on Preservation Scenarios for Projects Producing Digital Resources in the Area of Digital Humanities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creator</td>
<td>Dobreva, M.. (author)</td>
</tr>
<tr>
<td>Subject</td>
<td>Information Technology; Science; Technology--Philosophy; Computer Science; Digital Preservation; Digital Records; Science and the Humanities.</td>
</tr>
<tr>
<td>Description</td>
<td>Scenarios which address the digital preservation specifics in the three cases which will help readers to understand better how to take digital preservation related decisions in their own projects.</td>
</tr>
<tr>
<td>Publisher</td>
<td>HATII, University of Glasgow; University of Edinburgh; UKOLN, University of Bath.</td>
</tr>
<tr>
<td>Contributor</td>
<td>Joy Davidson (editor)</td>
</tr>
<tr>
<td>Date</td>
<td>12 November 2013 (creation)</td>
</tr>
<tr>
<td>Type</td>
<td>Text</td>
</tr>
<tr>
<td>Format</td>
<td>Adobe Portable Document Format v.1.3</td>
</tr>
<tr>
<td>Resource Identifier</td>
<td>ISSN 1747-1524</td>
</tr>
<tr>
<td>Language</td>
<td>English</td>
</tr>
<tr>
<td>Rights</td>
<td>© HATII, University of Glasgow</td>
</tr>
</tbody>
</table>

Citation Guidelines

About the DCC

The JISC-funded Digital Curation Centre (DCC) provides a focus on research into digital curation expertise and best practice for the storage, management and preservation of digital information to enable its use and reuse over time. The project represents a collaboration between the University of Edinburgh, the University of Glasgow through HATII, UKOLN at the University of Bath, and the Council of the Central Laboratory of the Research Councils (CCLRC). The DCC relies heavily on active participation and feedback from all stakeholder communities. For more information, please visit www.dcc.ac.uk. The DCC is not itself a data repository, nor does it attempt to impose policies and practices of one branch of scholarship upon another. Rather, based on insight from a vibrant research programme that addresses wider issues of data curation and long-term preservation, it will develop and offer programmes of outreach and practical services to assist those who face digital curation challenges. It also seeks to complement and contribute towards the efforts of related organisations, rather than duplicate services.

- Joy Davidson, Associate Director, Digital Curation Centre (UK)
- Kevin Ashley, Director, Digital Curation Centre (UK)

- Seamus Ross, Director, HATII, University of Glasgow (UK)
- Michael Day, Research Officer, UKOLN, University of Bath (UK)

Digital Curation Reference Manual copy editor
- Florance Kennedy, Administrator, Digital Curation Centre

Peer review board members have included
- Neil Beagrie, JISC/British Library Partnership Manager (UK)
- Georg Büechler, Digital Preservation Specialist, Coordination Agency for the Long-term Preservation of Digital Files (Switzerland)
- Filip Boudrez, Researcher DAVID, City Archives of Antwerp (Belgium)
- Andrew Charlesworth, Senior Research Fellow in IT and Law, University of Bristol (UK)
- Robin L. Dale, Program Manager, RLG Member Programs and Initiatives, Research Libraries Group (USA)
- Wendy Duff, Associate Professor, Faculty of Information Studies, University of Toronto (Canada)
- Peter Dukes, Strategy and Liaison Manager, Infections & Immunity Section, Research Management Group, Medical Research Council (UK)
- Terry Eastwood, Professor, School of Library, Archival and Information Studies, University of British Columbia (Canada)
- Julie Esanu, Program Officer, U.S. National Committee for CODATA, National Academy of Sciences (USA)
- Paul Fiander, Head of BBC Information and Archives, BBC (UK)
- Luigi Fusco, Senior Advisor for Earth Observation Department, European Space Agency (Italy)
- Norman Gray, Researcher, Department of Physics and Astronomy, University of Glasgow
- Hans Hofman, Director, Erpanet; Senior Advisor, Nationaal Archief van Nederland (Netherlands)
- Falk Huettmann, PhD Associate Professor, University of Alaska
- Max Kaiser, Coordinator of Research and Development, Australian National Library (Australia)
- Gareth Knight, Preservation Officer, CeRch, Kings College London
- Carl Lagoze, Senior Research Associate, Cornell University (USA)
- Nancy McGovern, Associate Director, IRIS Research Department, Cornell University (USA)
- Jen Mitcham, Curatorial Officer, Archaeology Data Service, University of York
- Mary Molinaro, Director, Preservation and Digital Programs, University of Kentucky Libraries
- Reagan Moore, Associate Director, Data-Intensive Computing, San Diego Supercomputer Center (USA)
- Sheila Morrissey, Senior Research Developer, ITHAKA
- Alan Murdock, Head of Records Management Centre, European Investment Bank (Luxembourg)
- Julian Richards, Director, Archaeology Data Service, University of York (UK)
- Donald Sawyer, Interim Head, National Space Science Data Center, NASA/GSFC (USA)
- Jean-Pierre Teil, Head of Constance Program, Archives nationales de France (France)
- Mark Thorley, NERC Data Management Coordinator, Natural Environment Research Council (UK)
- Helen Tibbo, Professor, School of Information and Library Science, University of North Carolina (USA)
- Malcolm Todd, Head of Standards, Digital Records Management, The National Archives (UK)
- Andrew Wilson, Senior Data Policy Advisor, Australian National Data Service
- Erica Yang, STFC Rutherford Appleton Laboratory
Preface

The Digital Curation Centre (DCC) develops and shares expertise in digital curation and makes accessible best practices in the creation, management, and preservation of digital information to enable its use and reuse over time. Among its key objectives is the development and maintenance of a world-class digital curation manual. The DCC Digital Curation Reference Manual (formerly the Digital Curation Manual) is a community-driven resource—from the selection of topics for inclusion through to peer review. The Manual is accessible from the DCC web site (http://www.dcc.ac.uk/resources/curation-reference-manual).

Digital Curation Reference Manual instalments provide detailed and practical information aimed at digital curation practitioners. They are designed to assist data creators, curators and reusers to better understand and address the challenges they face and to fulfil the roles they play in creating, managing, and preserving digital information over time. Each instalment will place the topic on which it is focused in the context of digital curation by providing an introduction to the subject, case studies, and guidelines for best practice(s). To ensure that this manual reflects new developments, discoveries, and emerging practices authors will have a chance to update their contributions annually.

To ensure that the manual is of the highest quality, the DCC has assembled a peer review panel including a wide range of international experts in the field of digital curation to review each of its instalments and to identify newer areas that should be covered. The list of current and previous members of the peer review board is provided at the beginning of this document.

The DCC actively seeks suggestions for new topics and suggestions or feedback on completed instalments. Both may be sent to the editors of the DCC Digital Curation Reference Manual at info@dcc.ac.uk.

Joy Davidson and Kevin Ashley
Digital Curation Centre

18 April 2011
Biography of the author

Dr. Milena Dobreva is a Senior Lecturer in Library, Information and Archive Sciences at the Faculty for Media and Knowledge Sciences, University of Malta. She was the principal investigator of EC, JISC and UNESCO funded projects in the areas of user experiences, digitisation and digital preservation and is a regular project evaluator for the EC. From 1990-2007 she worked at the Bulgarian Academy of Sciences where she earned her PhD in Informatics and served as the founding head of the first Digitisation Centre in Bulgaria. She was also a chair of the Bulgarian national committee of the Memory of the World programme of UNESCO. From 2007-2011 she worked for the University of Glasgow and the University of Strathclyde. Milena was awarded an honorary medal for her contribution to the development of the relationships between Bulgaria and UNESCO (2006) and an Academic Award for young researchers (Bulgarian Academy of Sciences, 1998).
Introduction and scope

Many large institutions in the cultural heritage sector wish to incorporate digital preservation activities into their everyday working practices and are seeking advice on how best this should be done. At the same time, multiple small-scale projects which produce digital resources – following diverse guidelines and standards – do not consider long term preservation at all in their project plans. The longevity of these projects’ outputs is inadequately addressed. It is not surprising then, that the sustainability of resources produced by digital humanities projects is a topic of discussion in recent literature (cf. Cantara 2006; Warwick et al. 2009).

There is ongoing reassessment of the responsibilities of those stakeholders who should be responsible for sustaining digital resources. As Linda Cantara notes:

“Traditionally, it has been the role of librarians, curators and archivists to ensure long-term viability of and access to cultural heritage materials… In contrast, digital preservation is characterized as a life cycle process – that is, ongoing management responsibility beginning with the creation of the resource and continuing for an indefinite duration.” (Cantara, 2006, p. 38).

This shift requires redefining the roles of the professionals who create and manage digital resources. The researchers involved in digital humanities projects constitute a new group of producers of digital resources. They are primarily concerned with the creation of the resources and with their quality and do not necessarily have the skills to apply preservation policies, workflows and tools. In addition to this, although multiple initiatives and projects address digital preservation issues, it is not easy to find solutions that can be applied in every instance.

Recommendations for the sustainability of results are often quite generic, for example: “The ideal digital resource should keep documentation and make it available from the project website, making clear the extent, provenance and selection methods of materials for the resource.” (cf. Warwick et al. 2007, p. 300). Abby Smith provides a good introduction to preservation within the digital humanities (cf. Smith (2004)), but again it addresses commonalities and does not provide details on the practical implications for ongoing projects.

More specific guidance for creators of digital resources is needed to systematically address preservation issues within digital humanities projects.

---

1 The British Library, the German National Library (DNB) and the Royal Library in the Netherlands are just two examples of institutions which apply successfully digital preservation solutions. For further information see Digital versus print as a preservation format – Expert views from international comparator libraries, Available on [http://www.bl.uk/aboutus/stratpolprog/ccare/introduction/digital/digitalvprint/index.html](http://www.bl.uk/aboutus/stratpolprog/ccare/introduction/digital/digitalvprint/index.html). DNB takes part in the SHAMAN project and is applying its demonstrator as part of its activities.
In this chapter we concentrate on long term preservation as only one aspect of sustainability. It is however, one of the most vital, since without the preservation of the digital objects, we cannot speak of any other aspects of sustainability. Sometimes long-term preservation is seen to be complete once a digital copy of the physical object has been produced. In reality, many other aspects are involved. One popular definition of digital preservation, used within the SHAMAN project, is that digital preservation is “interoperability into the future”.\(^2\) Such future-proofing is a common concern for a wide range of digital content producers and providers.

How can small projects – especially in the cultural heritage and digital humanities domains – address making their outputs interoperable into the future? What do they need to consider and what elements of the project methodology can help guarantee not only that digital objects remain physically intact in the future, but that they are also usable? To ensure future use and understanding of the object, electronic resources must be accompanied by preservation quality metadata.

In this chapter we will present various scenarios and recommendations for smaller-scale projects especially in the area of digital humanities. In the Methodology section we will provide a brief overview of digital preservation and its current standing. We will use digital preservation terminology as suggested by ISO 14721:2003 (Space data and information transfer systems – Open archival information system – Reference model), widely known as OAIS (see OAIS 2003). We will then take as case studies three projects – TEXTE, KT-DigiCult-Bg and the Parallel Archive – and synthesise scenarios drawn from these which describe different challenges and approaches to long-term preservation.

\(^2\) SHAMAN (Sustaining Heritage Access Through Multivalent ArchiviNg, project funded by FP7, [http://www.shaman-ip.eu/](http://www.shaman-ip.eu/))
Background and developments to date

The evaluation on the Progress on the digitisation and online accessibility of cultural material and digital preservation across the EU, published in August 2008 by the European Commission states:

“The absence of clear and comprehensive policies in many Member States was identified in the Recommendation as a threat to the survival of digitised and born-digital material…”

It also provides a good practice example from the UK, the Digital Preservation Coalition. Although this British experience is given as an example of a successful approach, the implementation of preservation in the current digital repositories in the UK is far from being satisfactory. The recent final report of the DRAI project (see Daisy Abbott 2008) on UK-based digital repositories identified that

“Of the 3,707 collections aggregated, only 737 (20%) had evidence of a preservation policy, and of these the vast majority were from within the Arts and Humanities Data Service (AHDS) Collections Repository which, since March 2008, is no longer funded, leaving only 6% of resources with ongoing policies for long term preservation.”

With digital preservation recognised as being of high importance, why is it still not implemented widely? The Blue Ribbon Task Force Report (2008) summarized five major obstacles to building sustainable digital preservation programs:

- inadequacy of funding models to address long-term access and preservation needs;
- confusion and/or lack of alignment between stakeholders, roles, and responsibilities with respect to digital access and preservation;
- inadequate institutional, enterprise, and/or community incentives to support the collaboration needed to reinforce sustainable economic models;
- complacency leading to current practices not being good enough; and
- fear that digital access and preservation is too big to take on.

While these mostly address stakeholder-related issues, the DPE report of 2007 reviewed the progress of digital preservation research world-wide. It concludes that “The analysis of the last 16 years of effort in the area of digital preservation… support[s] our claim that while much work has been done on

---

3 Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions Europe’s cultural heritage at the click of a mouse. Progress on the digitisation and online accessibility of cultural material and digital preservation across the EU [SEC(08) 2372] of 11 August 2008. Available:
the periphery, the central problems of digital preservation have yet to be resolved.” Automation in digital preservation is one of the ten areas named by this study as being in a need of accelerated research.

These two reports identify two key areas which need attention: the organisational aspects of, and the research into, preservation. Any institution which currently faces the preservation challenge has to make difficult choices within the context of a rapidly changing technological environment and the lack of a generally recommended and effective solution.

The deeper reason why, after two decades of efforts, digital preservation has still not reached maturity, is the lack of a consistent theoretical foundation. In 2001, James Cheney et al. articulated the need to develop a mathematical theory for the area of information preservation. Such a theory would strengthen understanding of the object models and basic transformations which need to be supported in a preservation system; it will also be of great benefit to automation because it will make it clearer which processes can be automated. In addition, a coherent theory would help those working in digitisation to identify gaps in their existing procedures and data flows, contributing to the implementation of systems whose behaviour could be consistently traced. However, since this first suggestion for a theory of preservation was made, there has been little progress made in developing one. This is partly due to the fact that identifying the basic elements needed for such a theory is not a trivial task.

In addition to this methodological complexity, the variety of domains where digital preservation is to be applied brings further difficulties. It is not clear whether a universal ‘one-size-fits-all’ solution could be used across different domains. For example, in the cultural heritage domain considered in this paper, there are various curatorial situations within which digital objects are located. Resources prepared by small projects are in more danger of lacking longevity compared to those of large institutions – particularly in cases where the lifecycle for digital object curation is not applied in full. It is not our aim in this paper to investigate the levels of awareness of digital preservation issues across the whole range of cultural heritage institutions – rather, we highlight the current state of affairs within the field more generally, assessing which issues need to be addressed by projects in the cultural heritage domain.

In 2002, the Reference Model for an Open Archival Information System (OAIS) was published and one year later adopted as an international standard ISO 14721:2003, see (OAIS: 2003). It provides a functional approach helpful for understanding and describing the processes taking place within a preservation system. In addition its information model helps to understand the changes in the digital object within the archival system. However, the specification of the minimum information which should accompany a digital object in order to guarantee its accessibility, interpretability and usability in the future, is not addressed.

There are various ongoing attempts to build a theory of preservation. In 2007, Giorgos Flouris and Carlo Meghini again suggested basing such a theory on
firm mathematical foundations and this was addressed within the CASPAR\textsuperscript{4} project. In the same year, Paul Watry presented the future directions of research of the SHAMAN\textsuperscript{5} project. The project later addressed the issue of preservation context, which needs to be stored within the archival system in order to guarantee that the essential properties of digital objects (such as authenticity and integrity) will be maintained over time. Reagan Moore and MacKenzie Smith suggested a practical approach, demonstrating how various preservation system requirements can be implemented as rule sets (Moore & Smith 2007).

In the next section we summarise the preservation-related issues which need further development in the context of digital humanities.

\textsuperscript{5} SHAMAN – Sustaining Heritage Access through Multivalent ArchiviNg an Integrated Project co-financed by the EU within FP7. http://shaman-ip.eu/shaman/
How the topic applies to Digital Curation

The preservation of digital resources created within small projects inevitably calls for clear answers to two questions:

- Which digital preservation solutions are best fitted to the cultural heritage domain?
- How should these be applied (including, how to plan for preservation within the project lifecycle)?

The dependence of preservation on metadata is a further area of concern because the evaluation of metadata across various types of digital resource shows they are not consistently applied. For example, Zhang and Jastram (2006) analysed the differences in metadata supplied by various professional communities in their study of human behaviour in metadata creation. They compiled a sample of 2400 web sites from four professional groups – each represented by 600 web sites – in order to ascertain how many of each contained some form of metadata. They found that 51.17% of the websites created within the Library and Information Science community contained metadata; compared to 66.5% of those from the Information Technology community, 66.7% of those created by government and not-for-profit organisations, and 67% of those created by the Business and Industries sector. It is alarming that almost half of the websites within the Library and Information Science community do not contain metadata (even more because this community should be better prepared professionally to make intelligent use of metadata).

Even in the cultural heritage domain, where the value of metadata is well-understood, there are problems with its application, particularly with regard to standards and formats. For example, Warwick et al. (2009) found that “In the absence of technical documentation, it was impossible to reuse text files with inconsistent mark-up (COCOA and XML) in a Digital Library. Also, although users require procedural documentation, about the status and completeness of sources, and selection methods, this is often difficult to locate”. This article analyses the use of digital materials, however the same issues would apply to long-term preservation: on the one hand because inconsistent metadata supplied with objects cannot be effectively exploited or utilised, and on the other, because the authenticity of the objects will not be traceable.

Another worrying example comes from the evaluation of a German national digitisation programme given as an example at a brainstorming session organised by the DELOS project in 2005. This revealed that “insufficient metadata practice, endanger[s] the usage of the digital documents, not to speak of their preservation: 33% of the objects had no metadata at all, 33% bibliographic metadata only, 10% had both bibliographic and subject metadata (and the rest had no information). Less than a third of the metadata was digital.” It is difficult to gather subsequent evidence on the changes in

---

6 DELOS brainstorming on Metadata topics, Juan les Pins, 05/12/2005, http://www.ukoln.ac.uk/ukoln/staff/t.koch/pres/Brainst200512-MDc.html
metadata practice – but again it is alarming to find estimates demonstrating such neglect of metadata.

These examples illustrate how serious the metadata problem is, but the inclusion of proper documentation and metadata recommended by Warwick et al. (2007) only solves one aspect of the long-term preservation issue. Another more fundamental problem is what steps a project which produces digital resources should take in order to guarantee that the digital objects will not be lost.

In response to the increased need for the introduction of common practices in assuring long term digital preservation (DP) of digital objects, the International Standards Organisation (ISO) have, in the last decade, developed a number of conceptual DP standards and also some technical guidelines. The most popular standard in this area is ISO 14721:2003 (Space data and information transfer systems – Open archival information system – Reference model), widely known as OAIS. It is a conceptual framework which identifies the main functional components and basic data flows within a digital preservation system. The development of OAIS arose from the need for a model which would specify the basic components of a system for long-term preservation of digital objects and their relationships with the ‘external world’.

In Fig. 1, an OAIS is represented as a box which is connected to other entities from the external environment. These are Producers, Consumers, and Management (the model does not suggest what roles might be defined within the archive). A special class of Consumer is the Designated Community. The Designated Community is the set of Consumers who should be able to understand the preserved information.

![Figure 1: The OAIS functional entities and environment](source_image)

---

7 A webpage which provides information on the basic stages of its development is [http://nost.gsfc.nasa.gov/isoas/](http://nost.gsfc.nasa.gov/isoas/).
8 For a record of the activities see [http://nssdc.gsfc.nasa.gov/nost/isoas/](http://nssdc.gsfc.nasa.gov/nost/isoas/).
10 Source: Figure 4-1, page 4-1 (OAIS, 2003).
Understanding the actual needs and the level of background knowledge of the designated community is non-trivial in the cultural heritage domain, especially when we consider members of that community who are creating scholarly resources. Although digital cultural heritage is discussed as one domain, it brings together multiple specialised domains—e.g. manuscript studies, numismatics, epigraphy, art studies, theology. The knowledge of specialists from different domains differs—as well as the needs and the expectations of those specialists. We must pay special attention to the concept of "designated" community as it is essential for understanding those resources both currently and in the future.

Alongside this, we might consider another important issue: the dynamics of resources. While digital resources created by memory institutions such as libraries, archives and museums are generally ‘static’ because they represent the fixed, physical resources of the institutions, those created by projects tend to be dynamic as they could be enriched and extended as the project develops. Such resources do not usually cover a fixed set of holdings but are grouped according to a specific theme, with only a selection of holdings chosen for digital representation. After a core set of resources has been made available, more may be added at a later stage. Being familiar with the model will help cultural heritage professionals to better organize the digital objects’ lifecycles and take informed decisions on the preservation aspects of their resources.

In addition to the definition of functional entities, an OAIS information model explains the data flows both within the archive and between the environment and the archive.

Every submission of a piece of information to an OAIS by a Producer, and every instance of dissemination of a piece of information to a Consumer, can occur either as a single discrete transmission, or as a series of transmissions. To describe the exchange of data, OAIS defines the concept of an Information Package as a “container” of two types of information: Content Information (CI) and Preservation Description Information (PDI). The CI and PDI are viewed as being encapsulated and identifiable by the Packaging Information (PI).

OAIS defines three specialised types of information packages (IP):

- Archival Information Package (AIP): An Information Package, consisting of the Content Information and the associated PDI, which is preserved within an OAIS.
- Dissemination Information Package (DIP): The Information Package, derived from one or more AIPs, received by the Consumer in response to a request to the OAIS.
- Submission Information Package (SIP): An Information Package that is delivered by the Producer to the OAIS for use in the construction of one or more AIPs.

PDI is divided into four types of preservation information: Provenance, Context, Reference, and Fixity. Provenance describes the source of the Content Information; Context describes how the CI relates to other
information outside the Information Package. Reference provides identifiers, by which the CI may be uniquely identified. Fixity provides a wrapper, which protects the CI from undocumented alteration. A key point to note is that OAIS does not suggest any specific metadata as instantiations of PDI. There is therefore no guidance on what constitutes a minimum technical requirement in respect of representing and encoding PDI information within corresponding PDI data bitstreams. The definition of a minimum required set of data should be based on a study of what is necessary to assure a reliable, consistent, and measurable specification and implementation of a preservation system.

The consequence for the cultural heritage domain is that currently there is no single and uniformly accepted set of metadata elements which guarantee the longevity of resources. If an institution or a project plan for preservation (which in fact needs to be included from the beginning of the project), they still face the need to decide how to structure the different types of archival packages and what metadata need to be stored for preservation purposes.

In OAIS the closest analogue to the high-level notion of metadata can be found in the idea of representation information, while on the implementation level, PDI provides the closest notion to a particular instantiation of a set of metadata elements.

Having such a reference framework for long-term digital preservation systems has made an essential contribution to the development of a common professional understanding and vocabulary. It has had a significant impact on the development of standards and has been an effective guide for those setting up preservation systems. Yet, the complexity of the area and the flexibility of the model allows for multiple interpretations of how the OAIS framework can be implemented in real life applications. This arguably gives rise to a need to develop implementation guidelines and practical policy frameworks.

**Metadata for Preservation vs. Preservation of Existing Metadata**

There are two key issues which need to be considered concerning metadata and preservation in the cultural heritage domain:

- Identifying what metadata is needed for preservation purposes. Besides supporting a reliable preservation process, it should help make resources understandable to their designated communities.
- How to preserve the metadata accompanying existing digital objects. Since it is often the case that cultural heritage resources already have extensive metadata, with multiple schemes in use, this issue also touches upon the understanding and use of existing metadata schemes in the future. The interoperability of metadata schemes in general is essential – but here we consider it as it relates to time – i.e. interoperability between now and the future.

In the area of preservation metadata, the basic concern, as stated by Brian Lavoie and Richard Gartner, still seems to be the development of schemes which define preservation elements (2005). Presentation and interoperability
issues are not the only concerns: preservation metadata, as all other types of metadata, are affected by the “metadata bottleneck”\textsuperscript{11}, a metaphor which indicates that the human efforts needed to create metadata can not keep pace with the creation of new digital resources.

In 2002 OCLC (Online Computer Library Center, Inc.) and RLG (Research Libraries Group) created a working group to explore how a consensus between different projects and initiatives in the field of preservation metadata could be achieved, with the aim of devising an OAIS-compliant preservation metadata solution. A body of recommended metadata for digital preservation was developed in the following years under the name PREMIS: PREservation Metadata Implementation Strategies (see PREMIS (2008)).

Attempting to define clearly preservation metadata presents many challenges. What preservation metadata should be used? What minimum set needs to be supplied in order to guarantee a reliable preservation process? How can the creation of preservation metadata be automated? Can it be guaranteed that the digital resources developed within a particular project are always accompanied by sufficient preservation quality metadata? And can there be interoperability between multiple existing schemes?

In the last few years, several metadata schemes have been developed dedicated to long-term preservation. In the cultural heritage domain these new metadata types have to be combined with existing descriptive metadata based on traditional cataloguing schemes. Amongst those new types developed are The Metadata Encoding and Transmission Standard\textsuperscript{12} (METS), a container format for metadata and content files maintained by the Library of Congress; and LMER (Long-term preservation Metadata for Electronic Resources)\textsuperscript{13} developed by the German National Library.

The practical preservation activities required to combine several metadata schemes to achieve a reliable preservation process is a topic that is currently being considered by many researchers. Angela Dappert and Markus Enders (2008) presented an example of integrating METS, PREMIS, and MODS\textsuperscript{14} metadata in an OAIS-compatible system which aims to ingest, store, and preserve ejournals in the British Library. They demonstrated that no single existing metadata schema accommodates the representation of descriptive, preservation and structural metadata necessary for such a system.

In the context of the Digital Curation Centre (DCC) lifecycle, the content of this chapter is relevant to

- data creation / data producers, because they need to choose appropriate digital object models which would address the needs of digital preservation – something which still is not well developed in the digital humanities domain.

\textsuperscript{11} The term metadata bottleneck was coined by E. Liddy in 2002.
\textsuperscript{12} http://www.loc.gov/standards/mets/
\textsuperscript{13} LMER description and LMER schema: http://www.d-nb.de/eng/standards/lmer/lmer.htm
\textsuperscript{14} http://www.tei-c.org/index.xml
- data curation / data curators, because they need to understand the specifics of objects created by different projects and the wide variety of practices and metadata schemes in the digital humanities domain.
- data reuse / data consumers, who should be able to use the objects (if not properly documented the future use would be compromised).

This chapter provides a general analysis of basic digitisation project concerns; however, when we address cultural heritage resources it seems that digital preservation issues extend far beyond the basic. The digital resource preparation stage of such projects is particularly resource-intensive, and caring for project outputs might seem to be outside the scope of the project teams' responsibilities. In fact, digital preservation is a domain in its own right and either the project team should include professionals in this area, which could be problematic in smaller-scale projects, or other team members need to find their own way vis-à-vis long-term preservation.
**Topic in action**

In this section we take the OAIS model as a starting point and analyse three sets of digital resources prepared in different environments. In order to cross-compare the resources the following list of threats as summarised in (PARSE.Insight : 9) has been used.

<table>
<thead>
<tr>
<th>Threat</th>
<th>Requirements for solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users may be unable to understand or use the data e.g. the semantics, format, processes or algorithms involved</td>
<td>Ability to create and maintain adequate Representation Information</td>
</tr>
<tr>
<td>Non-maintainability of essential hardware, software or support environment may make the information inaccessible</td>
<td>Ability to share information about the availability of hardware and software and their replacements/substitutes</td>
</tr>
<tr>
<td>The chain of evidence may be lost and there may be lack of certainty of provenance or authenticity</td>
<td>Ability to bring together evidence from diverse sources about the Authenticity of a digital object</td>
</tr>
<tr>
<td>Access and use restrictions may make it difficult to reuse data, or alternatively may not be respected in future</td>
<td>Ability to deal with Digital Rights correctly in a changing and evolving environment</td>
</tr>
<tr>
<td>Loss of ability to identify the location of data</td>
<td>An ID resolver which is really persistent</td>
</tr>
<tr>
<td>The current custodian of the data, whether an organisation or project, may cease to exist at some point in the future</td>
<td>Brokering of organisations to hold data and the ability to package together the information needed to transfer information between organisations ready for long term preservation</td>
</tr>
<tr>
<td>The ones we trust to look after the digital holdings may let us down</td>
<td>Certification process so that one can have confidence about whom to trust to preserve data holdings over the long term.</td>
</tr>
</tbody>
</table>

Table 1: Summary of threats and recommendations of PARSE.insight project (PARSE.Insight Deliverable D2.2. (2010), p. 8)

**The TEXTE Project**

TEXTE is an ongoing project coordinated by The Moore Institute based in the National University of Ireland (NUI), Galway. It combines four sub-projects which aim to illustrate how the current range of computer-enhanced editing tools can be used to produce high-quality scholarly editions (mediaeval manuscripts; correspondence; a set of journal issues and ephemera)\(^{15}\). The

materials are digitised and then transcribed and encoded using the Guidelines of the Text Encoding Initiative. The resulting resources are highly annotated scholarly editions, comprising a combination of the original texts and scholarly findings.

The project was headed by Prof. Sean Ryder, Chair of the Department of English at NUI Galway. Tim McLoughlin worked on a digital edition of the correspondence of the Irish neo-classical painter James Barry. He synthesized his experience on the differences of editing for traditional and digital editions in McLoughlin (2008). Malte Rehbein created a new, genetic form of digital edition with his work on the Kundige Bok, a late medieval manuscript containing the town statutes of Gottingen in Germany (cf. Rehbein 2009). John Moulden worked on the "Cleland collection", a group of eighteenth and early-nineteenth century popular songbooks and prose books that once belonged to a farming family in County Down, now held by the Ulster Museum. Paul Caton created an edition of the Dublin Penny Journal, a weekly magazine published in Dublin from June 30th, 1832 until June 25th, 1836.

Simultaneously, Francesca Benati and Justin Tonra were working on the Thomas Moore Hypermedia Archive in the Moore institute.

All these digital editions were built using the same guidelines for text encoding but the objects created for the individual digital editions are not identical in structure. Having a common framework facilitates the grouping together of these resources online for publication and visualisation, supporting the development of common, integrated ways of annotating resources from the same historical and cultural background.

From a digital preservation point of view, one issue needs to be studied further. The TEI P5: Guidelines for Electronic Text Encoding and Interchange do not suggest which metadata elements should be used specifically for preservation purposes. It is essential to have a well-formed <TEI.Header> and to document the principles applied in the encoding of the text; these are prepared as a part of the project documentation. But the <TEI.Header> does not currently support elements which can help to trace the digital object’s authenticity, integrity and chain of custody – these could be included as additions, but a difficulty for such projects is the lack of guidance on the significant properties of digital objects which need to be supported. Since significant properties vary depending on the specific material and the purpose for which they are intended it is difficult to have one recommended solution, but better guidance for digital projects in the Humanities is still much needed.

Projects such as TEXTE which develop textual resources would benefit from clear guidance on this matter; this project faced the need to find its own solution. The final product of TEXTE is a set of web resources; their inclusion into a repository is still under question; and resources are maintained within

---

the Institute hosting the project while some of the staff members believe that the task of a storage and provision of access should be with the University Library.

Given that creating these digital editions already requires extensive mark-up, it would be easy for staff members to extend the TEI encoding practice to supply necessary preservation metadata. Yet, one could also justify an approach whereby the preservation metadata would not be included with the digital objects themselves, but in the submission information packages supplied when digital objects are first ingested into an archive. More research is needed to determine which approach is best suited to such projects.

If in future such digital resources are to be ingested into a repository, some additional effort will be needed to define submission information packages (SIPs). Should such a step be taken, this would happen after project staff finished their work on the project, since the project was a FP6\(^{19}\)-funded transfer of knowledge, and four staff members were incoming researchers for a fixed time. However, taking care of such issues after a project has finished is not the best possible approach – it should become a standard practice to address future preservation from the inception of projects. Otherwise the risk of actual neglect of preservation is very high, especially when there is no financial arrangement for this activity and the funding of the project is already over.

Regarding the other metadata-related issue – the longevity of specific metadata schemata applied in the cultural heritage domain – the decision to use a widespread encoding framework seems to guarantee (as far as is possible) that the digital objects will be usable in the future. However, even if a project uses an encoding which is common to the larger professional community, the only guarantee that future generations will be able to properly use the digital objects is if this community can ensure that its knowledge base (in its OAIS sense) will be known to the future users.

In a series of interviews with the project staff held between June and August 2009, a number of concerns were revealed:

- Most participants were reluctant to introduce additional elements to TEI P5 because this would create difficulties for interchange; they also recognised that this would need to be documented and would constitute an additional layer of complexity. Some of them were nevertheless convinced that additional mark-up elements for their specific types of materials would be of benefit.
- The editions contain processed images from scanned materials. It was not clear who should store the original scans as master copies; again the expectation was that this should be the library but apparently such a policy was not in place.
- The individual digital editions employ different delivery mechanisms; these delivery mechanisms vary in composition and

\(^{19}\) Framework Programme 6 of the European Commission.
complexity. For example one of the researchers used XTF\textsuperscript{20}, a free indexing, display and query tool for heterogeneous collections, but the implications of any changes of the software tools in the future with regard to this component of the delivery mechanism were not considered. Another team member created his own software (using \textit{java servlets}) for the edition he was working on. In this case subsequent preservation would require also appropriate software documentation (and this would be an issue for many projects designing their own specialised tools).

- Generally the researchers were happy with the developed functionality of the digital editions but also typically wanted to be able to add more options which would be tailored to their particular needs.
- The researchers shared the opinion that the final version is the one which needs to be sustained.
- One team member suggested that dedicated staff administering digital projects should be based in the library; this would mean creating new job roles within the library permanently to administer such resources. It would also have implications on an official deposit policy for the institution.

What are the implications of these observations for preservation? In this particular project all team members used the same metadata scheme (TEI P5); if this is sustained in the long-term, their materials will be usable in perpetuity. Researchers see the library as the obvious choice for ensuring the sustainability of digital resources, but institutionally such policies are not in place. Preservation issues relating to the specific functionalities offered and tools used to create and annotate the resources were not considered. This is however an important issue, especially for this type of resource. Since the researcher team expect that only the final version of their editions needs to be sustained, this would mean re-ingest of the complete editions into an archive after older versions and their respective objects had been removed. Such a scenario would have implications for other resources which referred to these older versions – again, there is not space to consider this here.

The KT-DigiCult-BG Project

The KT-DigiCult-BG project funded by the Marie Curie programme of FP6 reached completion in 2008. One of its outcomes was the creation of the Digitisation Centre in the Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences (IMI-BAS). This centre digitised a substantial number of various types of Bulgarian cultural and scientific heritage objects (see Table 2).

Two very clear issues identified early on in this project were the organisation and long-term preservation of these objects. The selected objects were not deeply annotated and are intended for use by the general citizenry. The metadata associated with these various archival materials – old printed books,

photographs and periodicals – followed various encoding standards and were not encoded like the metadata of the TEXTE project.

The heterogeneity of this approach leads to a different situation in the produced resources. Currently the digitisation centre team is planning to place all scientific resources into a DSpace repository\textsuperscript{21} which will enable the use of its resources within the DRIVER\textsuperscript{22} project's repository, funded by another EC project, OpenAIRE.\textsuperscript{23}

These resources will be transformed according to this larger initiative's requirements and the relevant SIP will be constructed. In this scenario the preservation task (with regard to the digitised documents prepared for access in PDF format) stays with the repository while the digitisation centre will take care of the physical copies of the master files, which are stored in TIFF format. Table 2 below shows what mixture of materials had to be considered:

<table>
<thead>
<tr>
<th>Type of material</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old printed books (National Library Ivan Vazov, Plovdiv)</td>
<td>17,000 pages</td>
</tr>
<tr>
<td>State Archives</td>
<td>8,000 pages</td>
</tr>
<tr>
<td>Research Archive of the Academy of Sciences</td>
<td>24 archival files, 1,000 pages</td>
</tr>
<tr>
<td>Old periodicals (newspapers)</td>
<td>1,200 pages</td>
</tr>
<tr>
<td>Musical periodicals</td>
<td>1,000 pages</td>
</tr>
<tr>
<td>Archive of the Higher Attestation Commission</td>
<td>8,500 pages</td>
</tr>
<tr>
<td>Archive of IMI-BAS</td>
<td>18,000 pages</td>
</tr>
<tr>
<td>Mathematical heritage (the heritage of Prof. Obreshkov)</td>
<td>2,000 pages</td>
</tr>
<tr>
<td>Audio archive of Bulgarian dialects</td>
<td>7 hours</td>
</tr>
<tr>
<td>Mathematical periodicals (1905 - 1991)</td>
<td>29,000 pages</td>
</tr>
<tr>
<td>Old photographs</td>
<td>200 photographs</td>
</tr>
<tr>
<td>TOTAL</td>
<td>85,900 units</td>
</tr>
</tbody>
</table>

Table 2: Digitised Resources in 2005-2008 in IMI-BAS

Consultations with the members of the Digitisation Centre held in March 2010 revealed that they store their master files locally. They also have a preservation plan for media refreshment (copying the digital objects to new media). The centre does not have a good internal document delivery system and has so far produced a number of isolated websites\textsuperscript{24} The preparation of materials for inclusion in a larger repository which follows clear guidelines on metadata structures is regarded as an issue of major importance. Preservation-wise, if the selected repository has a good digital preservation strategy this helps guarantee the sustainability of the materials prepared for access. The centre has not yet considered this issue in depth.

\textsuperscript{21} http://www.dspace.org/
\textsuperscript{22} DRIVER: Networking European Scientific Repositories, http://www.driver-repository.eu/
\textsuperscript{23} Open AIRE, Open Access Infrastructure http://www.openaire.eu/
\textsuperscript{24} For example the archive of the founder of the Bulgarian Academy of Science Marin Drinov, http://www.math.bas.bg/digi/drinov/ (in Bulgarian only).
In this case preservation is not seen as a basic internal task for the digital objects used for delivery, but only for the master files which store digitised objects.

**Parallel Archive**

The Parallel Archive\(^2\) is a dynamic environment for researchers who would like to share digital copies of the primary sources they are using. Based on the simple idea that the research community can put into circulation a large number of such sources, without being dependent on large digitisation programmes within memory institutions, its priority is clearly user driven. Researchers can share materials of interest and that they have actually used in their research. To support this idea, The Parallel Archive offers a range of social networking functions including tagging, forums and user-comments. This places the primary sources and their digital equivalents within the contexts which users find to be of interest. The documents in The Parallel Archive are stored as original image files; files in PDF format are generated for access purposes and metadata stored in a designated database.

The preservation issues which the Parallel Archive faces include how to preserve both the digital objects and the dynamic additional metadata added to objects over time by different users.

The Parallel Archive is not only aware of preservation issues but is actively seeking its own solutions. On the one hand, the objects which need to be preserved are in formats known in advance (including a range of image formats, PDF files and a custom database). Older versions of PDF presented some difficulties when processing the objects (e.g. not rendered properly by current PDF readers), but the archive is aware of this. A more serious issue is the handling of the social media-type data which are dynamic; this issue remains under discussion.

This case is of interest because it combines two types of digital objects. The first group, which includes images, PDF files and databases, is more trivial to address preservation-wise. But the combination of static with dynamic content in a preservation effort is currently a challenge.

**Discussion**

The three cases presented above illustrate different situations with regard to long term preservation. These differences are summarized in Table3.

\(^2\) [http://www.parallelarchive.org/](http://www.parallelarchive.org/)
<table>
<thead>
<tr>
<th>Project</th>
<th>Resource Type</th>
<th>Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEXTE</td>
<td>Stand alone digital works</td>
<td>Four sets of resources, 1 platform (TEI-based editions). Beyond simple digitisation – extensive scholarly editing. Not clear if the resources will be developed further. Expectations that the institutional library will store the master files while the final digital editions are stored within the research institute which worked on the project.</td>
</tr>
<tr>
<td>KT-DigiCult-BG</td>
<td>Materials intended for inclusion within a bigger repository</td>
<td>Growing number of digital objects, various types of materials, different formats and metadata. Master files preserved locally, the issue is the preservation of objects used for delivery.</td>
</tr>
<tr>
<td>Parallel Archive</td>
<td>Dynamic content with a social networking dimension</td>
<td>Digital objects which are ingested to the repository are being converted to PDF. In addition to the original images there is a custom database which needs to be preserved. The basic issue is the preservation of dynamic content.</td>
</tr>
</tbody>
</table>

Table 3: Differences between the three case studies

If we analyse the threats suggested by the PARSE.insight project we will find differences between the three projects (see Table 4).

<table>
<thead>
<tr>
<th>Threat</th>
<th>Implications for the three case studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users may be unable to understand or use the data e.g. the semantics, format, processes or algorithms involved</td>
<td>This is especially important in the case of TEXTE which relies on extensive scholarly mark-up. But this is alleviated given that TEXTE uses a well-known and documented metadata scheme. KT-DigiCult-BG is flexible in following the various metadata schemes of repositories which ingest the objects it created for delivery purposes. Parallel Archive needs to address social networking materials and their preservation.</td>
</tr>
<tr>
<td>Non-maintainability of essential hardware, software or support environment may make the information inaccessible</td>
<td>This is a threat in all cases; KT-DigiCult-BG may be seen as less threatened with the plan to ingest its outputs into a big digital repository.</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Threat</th>
<th>Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>The chain of evidence may be lost and there may be lack of certainty of provenance or authenticity</td>
<td>This is important for all the cases studies presented. However, it is of particular threat to the Parallel archive case study because of the ability it provides for multiple people to contribute and tag the documents. This could lead to possible provenance and authenticity issues.</td>
</tr>
<tr>
<td>Access and use restrictions may make it difficult to reuse data, or alternatively may not be respected in future</td>
<td>Again, this needs to be considered within all presented projects.</td>
</tr>
<tr>
<td>Loss of ability to identify the location of data</td>
<td>Especially threatening in the case of TEXTE because if its resources do not become part of a larger repository there should be clarity about who would deal with any subsequent changes of identifiers.</td>
</tr>
<tr>
<td>The current custodian of the data, whether an organisation or project, may cease to exist at some point in the future</td>
<td>This is a difficult issue for all presented case studies since KT-DigiCult-BG and TEXTE are finished. Governance models which will assign clear roles in the future of all projects are vital. TEXTE will also need contingency plans because resources not part of larger repository.</td>
</tr>
<tr>
<td>The ones we trust to look after the digital holdings may let us down</td>
<td>If a project ingests its resources to a repository as in the case of KT-DigiCult-BG, this is a serious issue. Certification is still under development in the area of digital preservation and this makes the trustworthiness of repositories harder to quantify.</td>
</tr>
</tbody>
</table>

Table 4: Summary of threats according to PARSE.insight project and how they impact on each of the case studies (PARSE.Insight Deliverable D2.2., 2010)
Next steps

The following scenarios emerge from the analysis of the three case studies.

Scenario 1
Preparation of stand-alone resources. Under this scenario, the set of digital resources is prepared as a stand-alone collection and does not necessarily become a part of a digital repository.

Such projects need to pay special attention to proper documentation of the decisions taken on the encoding applied within the project – this is not only necessary for sustainability, but also for reuse of resources in the future as demonstrated by Claire Warwick et al. (2009). Documentation which refers to a general standard like TEI or XML is not sufficient because it does not guarantee future understanding and interoperability.

A set of preservation metadata needs to be supplied for every single digital object. This is currently not a trivial requirement because there is no common view on the structure of preservation metadata. Again, the internal project decisions need to be well-documented.

Scenario 2
Preparation of resources which will be integrated into a larger digital repository. The following key considerations should be taken into account:

- It is essential to be familiar with the processes of ingest of digital objects into the repository.
- The structure of the SIPs for ingest should be discussed in advance, including the metadata requirements of the repository.
- The significant properties of digital objects need to be discussed with the digital repository. The repository should guarantee the retention and preservation of authenticity and chain of custody metadata related to the ingested objects.

Scenario 3
Preparation of a dynamic digital resource. The following key considerations should be taken into account:

- While work on preservation of separate dynamic resources, e.g. blogs, has started, the issue of a resource which combines static and dynamic objects needs further research.
- Such cases pose interesting issues for the concept of the knowledge base, and the preservation of context of objects. Presumably dynamic entries are created by people with different understandings of the knowledge base. The implications for the information packages’ structure and functionality have also to be addressed; in this context, ‘dynamic’ refers not only to the adding of content but also to changes in the knowledge base reflected by such additions.
The Analysis Matrix (see Table 5) summarises what issues need to be taken into account in these scenarios.

Table 5. Summary of digital preservation issues reflected in the three suggested scenarios

<table>
<thead>
<tr>
<th>Issue</th>
<th>Scenario 1. (stand-alone resources)</th>
<th>Scenario 2. (resources to be integrated into a digital repository)</th>
<th>Scenario 3. (dynamic resources)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital object</td>
<td>The digital object is prepared completely and solely within the project.</td>
<td>The digital object is likely to be enriched in order to be ingested into a repository. This enrichment can be done by various parties according to the adopted procedures for ingest into the repository.</td>
<td>The core digital object is static but it is connected to a number of dynamic objects.</td>
</tr>
<tr>
<td>Significant properties of the digital object</td>
<td>Small projects usually do not consider significant properties but they are essential as future evidence of authenticity, integrity and chain of custody.</td>
<td>The values of the significant properties which guarantee authenticity, integrity and chain of custody need to be supplied jointly with the digital object when it is being ingested into the digital repository.</td>
<td>Significant properties of dynamic objects need to be studied further.</td>
</tr>
<tr>
<td>Preservation actions</td>
<td>The preservation actions e.g. refreshment of media or migration to newer file formats will be applied within the host institution; it needs to plan for these otherwise the danger is that the collection as a whole (or separate objects) will be</td>
<td>The preservation actions are applied within the repository. The creators of the collection do not have to plan for this but need to make sure what the digital repositories policies are on media refreshment and format migration.</td>
<td>The preservation actions are applied within the institution which develops the dynamic resource.</td>
</tr>
<tr>
<td>OAIS implementation</td>
<td>It is unlikely that small projects will implement the complete set of OAIS functional entities.</td>
<td>The repository hosting the project outputs may implement a specific subset of OAIS functional entities; this should be discussed in advance as a sustainability guarantee.</td>
<td>OAIS needs to be adapted in the area of structures of information packages. Refinement of Ingest and Access functionalities should address the dynamic nature of the resource.</td>
</tr>
</tbody>
</table>

It is hoped that this succinct analysis will help projects to better understand the preservation-related options available to them.
Future developments

The three different case studies of recent or ongoing digitisation work presented in this chapter revealed quite different levels of awareness of preservation issues. It would be very helpful if an audit framework could be developed to address specific issues in the digital humanities domain. This might be similar to the Data Audit Framework (DAF) developed for the analysis of the data curation within Higher Education Institutions in the UK.

Such a framework would need to address several areas:

- **Policies** of the group which develops the resources (normally this is a specialised centre, department or institute but it may also be an individual researcher) and other institutional stakeholders (e.g. the Library); policies should include Intellectual Property Rights issues and analysis of risks.

- **Assuring asset continuity, including:**
  - Content within the resource – types, formats, volume.
  - Metadata management.
  - Technological components within the digital edition – creator, sustainability, conditions for use.
  - Dynamic components.

- **Resources, including:**
  - Analysis of staff efforts for sustainability of resources and sustainability of funding.

Such a framework would provide specific guidance to the creators of digital resources in the digital humanities domain.

Some practical recommendations for data producers include:

- Provide extensive documentation for your project, including structure of information packages (SIP and DIP in the case of cooperation with an external repository and also AIP in the case of local archiving).
- Clarify what preservation metadata could be included and document the used elements (while we do not have the high quality automated solutions).
- Be sure to provide information on provenance and authenticity for the separate digital objects at the time of ingest.
- For stand alone projects: choose popular, widely-used data formats and put in place a policy for media refreshment.
- For projects aimed at repository ingest: check the policy of the repository on digital preservation; decide how to treat the master files.
- For dynamic projects: there is no universal solution at the moment and areas to follow are web archiving and preservation of complex objects.
Conclusions

The analysis of digital preservation issues and the specifics of the digital humanities domain pinpoint several issues which are essential for digital humanities projects:

- The understanding of digital preservation standards and models is essential to the organisation of proper lifecycle management for cultural heritage resources; there are examples of solutions applied within large memory institutions but smaller projects and initiatives need to address these issues when they develop their sustainability plans.

- Widely used metadata standards in the cultural heritage domain, such as TEI, could offer better guidance on metadata element subsets designed especially to support long-term preservation. TEI provides mechanisms for tracking changes in the digital object; it also could be used to encode information related to other copies made such as master or preservation copies and access or use copies; and possibly to record any processes applied to a resource, but this is not sufficient because it reflects the process of creation of an object but this would not help to claim authenticity of this object. Currently there is no mechanism to check the integrity and the chain of custody of digital objects. However, TEI guidelines should suggest how to correctly represent these changes. For example Linda Cantara in 2005 suggested how to form an AIP for the resources of the Tibet Oral History Archive Project; more examples are needed in order to establish good practices which could be incorporated in other cases.

- OAIS does not suggest any specific metadata as instantiations of preservation description information and in particular in the cultural heritage domain there is no single and uniformly accepted set of elements which guarantee the longevity of resources; therefore it is necessary to define what metadata need to be stored for preservation purposes. Until such an element set is defined, individual projects need to work out a solution that is best for them and the needs of their designated communities.

We have presented scenarios on digital preservation derived from three case studies. This work shows that there is a clear current need to suggest good practices for such endeavours, because while the projects tackle the preservation issue, it is very difficult to find individual good models to guarantee the longevity of resources.

Digital preservation of digital humanities resources brings together several issues on various levels: institutional policies, standards and workflows, and

\[26\] The German National Library and the Royal Library in the Netherlands are examples of institutions which apply successfully digital preservation solutions.
practical implementation which also includes the use of metadata for preservation as well as the long-term preservation of domain-specific metadata.

The examples discussed help us to better understand how current projects are approaching long-term preservation. There is definitely an urgent need to suggest good practices for such endeavours; while the projects tackle the preservation issue, it is very difficult to find robust individual models to guarantee the longevity of resources. A framework which would provide practical guidance is seen as a tool to practically help stakeholders in this domain.
References


## Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIP</td>
<td>Archival Information Package</td>
</tr>
<tr>
<td>CASPAR</td>
<td>Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval</td>
</tr>
<tr>
<td>CI</td>
<td>Content Information</td>
</tr>
<tr>
<td>DIP</td>
<td>Dissemination Information Package</td>
</tr>
<tr>
<td>DP</td>
<td>Digital Preservation</td>
</tr>
<tr>
<td>DPE</td>
<td>DigitalPreservationEurope</td>
</tr>
<tr>
<td>DRIVER</td>
<td>Networking European Scientific Repositories</td>
</tr>
<tr>
<td>ISO</td>
<td>International Organisation for Standardisation</td>
</tr>
<tr>
<td>LMER</td>
<td>Long-term preservation Metadata for Electronic Resources</td>
</tr>
<tr>
<td>METS</td>
<td>Metadata Encoding and Transmission Standard</td>
</tr>
<tr>
<td>MODS</td>
<td>Metadata Object Description Schema,</td>
</tr>
<tr>
<td>OCLC</td>
<td>Online Computer Library Center, Inc,</td>
</tr>
<tr>
<td>PDI</td>
<td>Preservation Description Information</td>
</tr>
<tr>
<td>PI</td>
<td>Packaging Information</td>
</tr>
<tr>
<td>PREMIS</td>
<td>PREservation Metadata Implementation Strategies</td>
</tr>
<tr>
<td>RLG</td>
<td>Research Libraries Group</td>
</tr>
<tr>
<td>SHAMAN</td>
<td>Sustaining Heritage Access Through Multivalent Archiving</td>
</tr>
<tr>
<td>TEI</td>
<td>Text Encoding Initiative</td>
</tr>
<tr>
<td>XTF</td>
<td>eXtensible Text Framework</td>
</tr>
</tbody>
</table>
An annotated list of key external resources

**Data Audit Framework Development**
Provides a data audit framework designed for use within the British Higher Education Institutions.

**Data Audit Framework Pilot Projects**
Presents cases of applications of the Data Audit Framework to data collections across four British Higher Education institutions.

**Extending Data Curation to the Humanities: Curriculum Development and Recruiting**
http://cirss.lis.illinois.edu/pdf/DCEP_Humanities_IMLS_Proposal.pdf
Assessment of needs and curriculum approach for teaching data curation to the Humanities.

**Acknowledgements**
I would like to acknowledge the helpful discussion with the research fellows from the TEXTE project, mentioned in this chapter, as well as several colleagues who were of great help in discussing the issues presented in this chapter: Kathleen Menzies from CDLR of the University of Strathclyde; Assoc. Prof. Nikola Ikonomov from IMI-BAS with whom some early work in this area had been done jointly; and Gabriella Ivacs from the Open Society Archive in Budapest.