

# **Embedding a VRE in an Institutional Environment (EVIE)**

## **Workpackage 4: VRE Preservation Requirements Analysis**

### **1. Introduction**

- 1.1 In Workpackage 4, the EVIE project has analysed the digital preservation requirements specific to Virtual Research Environments (VREs). It has identified requirements to preserve collaborative intermediate research outputs, datasets and conversations, as well as documents. The metadata and format standards that might be applied have been explored. The workpackage has leveraged and extended existing and ongoing work within the British Library and that of its external partners with whom it is collaborating on digital preservation.
- 1.2 This report is the main deliverable of workpackage 4 and documents the requirements and likely solutions for digital preservation in a VRE. Key preservation issues are illustrated by discussing the generic VRE requirements and solutions within the context of the existing infrastructure at the University of Leeds. The document concludes with recommendations on the preservation of outputs from VREs.

### **2. Current status of preservation within virtual research environments**

- 2.1 The research notes and outputs of today's researcher are scattered across a wide range of paper and digital resources including email, personal hard-drives, instant messaging logs, and shared directories. Increasingly, outputs are starting to appear in institutional and discipline-based repositories, but this is often limited to the peer-reviewed version of a text document (the "postprint") and the other outputs such as experimental data, audio and conference presentations are still found distributed across a variety of platforms within the institution. As VREs provide increasing support for all stages of the research lifecycle, they also provide an opportunity to inject life into research outputs beyond the project. Ensuring research outputs are preserved ensures they are available for subsequent discovery and that the research communities' knowledge base grows.
- 2.2 VREs offer the chance to provide services and functionality that create an environment in which researchers can preserve their work for the long-term. Indications are that this is not happening at present within VRE projects<sup>1</sup>. Many are focusing on providing document management within a suite of collaborative tools. Such management may include the eventual storage of the document within a local repository, but there seems to be little concerted effort with regard to its long-term curation or access. Furthermore it was found that many involved in such VRE project work were not aware of the need for preservation action or the challenges that this work presented. With EVIE the intention is to go a step beyond local deposition and address the issue of the long term preservation of research outputs.

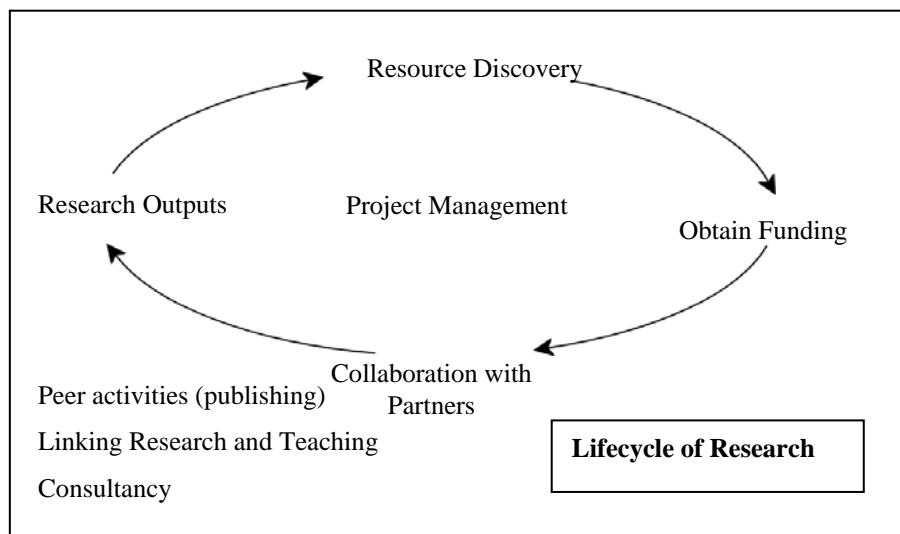
### **3. Why research outputs should be preserved**

- 3.1 The research lifecycle outlined in the EVIE Workpackage 2, User Requirements Analysis Report [1] shows how research outputs feed into the resource discovery

---

<sup>1</sup> Based on review of project documentation and discussions with VRE project staff in late 2005.

phase of research. Without such outputs the next phase of research cannot be built upon. The lifecycle diagram is reproduced here for information.



3.2 Whilst in many cases it is possible to discover these outputs via online aggregators of repositories, open access journals, and traditionally published journals shortly after they have been disseminated, this is not necessarily the case in the long-term. There is therefore a need to ensure that research outputs are preserved so that the discovery process is not interrupted and can be relied upon over time. Aggregation rarely includes data (which is becoming increasingly important to researchers) or additional multimedia information such as audio. Dissemination of research outputs is key to enhancing the standing of the institution and informing the Research Assessment Exercise, hence the importance of deposition within a recognised repository.

3.3 Preserving research outputs ensures that research outputs can be:

- Found
- Retrieved and accessed
- Understood and used
- Repurposed

## 4. What research outputs should be preserved?

4.1 There is already a mechanism for preserving peer-reviewed traditionally published material (e.g. subscription journals). In the UK this is centred on The British Library, which preserves copies of all UK published documents under legal deposit legislation. This legislation has recently been extended to cover digital material as well as print. The British Library also holds a vast collection of non-UK material to service research, again predominantly resulting from the standard publishing business model.

4.2 There is therefore a means by which published research outputs produced within the EVIE environment will appear in The British Library collections (either as print or digital formats) and therefore be included in the Library's Electronic Table Of Contents (ETOC). ETOC data is currently available via the Z39.50 Electronic Table Of Contents (ZETOC), Inside, British Library Direct and Google services as well as external hosts, thereby exposing the research through a number of discovery

channels and to a wide range of users worldwide. The British Library exposure is of course, only one route. Many journals are indexed by A&I services if the journal is deemed to be of sufficiently high quality, although these services do not tend to be as comprehensive as ETOC. UK books are also preserved within The British Library collection and are included in services such as COPAC, OCLC’s Find a Library as well as the institution’s own online catalogues.

4.3 Other types of output formats are not currently picked up by this mechanism. The Composite of outputs from research activity at University of Leeds (studied in EVIE WP3) identifies video, audio, software, data and presentations outputs as well as text which may pose significant concerns for preservation and access [2]. The outputs include a considerable range of content types, some of which might be considered for inclusion in the preservation workflow discussed below and others which may be addressed as part of the institution’s record management procedures. While the latter is outside the scope of this workpackage, the outputs to be addressed by a VRE preservation function represent a considerable challenge. A cross section of types of content is represented, although a detailed description of formats was unfortunately not available. Rusbridge provides a breakdown of file format categories [3], useful for analysis of these outputs, reproduced below:

- Media formats
- File formats created from hardware devices (e.g. digital cameras, scanners) and telemetry
- File formats created by programmers for specific projects
- File formats from standards-based, community or open source projects (perhaps not completely distinguishable from the previous case)
- File formats resulting from consumer-oriented commercial software products
- File formats from highly configurable products (e.g. SPSS)
- File formats protected by Digital Rights Management systems, or other forms of encryption or proprietary encoding.

Several of these categories are in evidence in the case studies. The table below provides some analysis of these categories, where they are evidenced as VRE outputs and the implications of this type of material with regard to enabling its preservation.

Format category	Evidenced in the Research Outputs Composite	Danger of impending obsolescence	Notes on the preservation implications
Media formats		Unknown	Digital preservation requires action to capture the significant properties and represent them as a bytestream (termed the Underlying Abstract Form by Cedars [4]). This process and the preservation of the resulting bytestream could vary in its complexity considerably depending on the type of content encountered.
File formats created from hardware devices (eg digital cameras, scanners etc) and telemetry	Raw data	High	Rapidly advancing developments both of commercial products and research technology can result in a lack of hardware (and software) support for useful data.
File formats created by programmers for specific projects and file formats from standards-based, community or open source projects	Test cases (in arbitrary formats), Processed data	Medium to high	Research outputs might consist of raw data captured in a particular experiment, unique software that provides analysis of the data, a set of parameters that will typically evolve over time, resulting data analysis output and documentation describing some or all of the above. All of these outputs, not just the data itself must be preserved (evidenced by: Source code, Documentation, XML Schemas, etc). Effective curation

			and preservation requires the capture of a complex set of interrelating Representation Information.
File formats resulting from consumer-oriented commercial software products	Photos, Video, Audio recording	Low	Perhaps the lowest risk category of the main VRE outputs. As Rusbridge points out, while these formats are not likely to disappear over night, a gradual slide towards obsolescence may occur. Tools from the Libraries and Archives community as well as the commercial world are being developed to address formats in these areas. Large scale European preservation initiatives like the PLANETS [5] and CASPAR [6] projects are notable examples that are likely to provide useful solutions in this area.
File formats from highly configurable products			Not unlike some of the problems associated with the classic eScience data, software, parameter, analysis and documentation problem described above. While the formats in this category may be better known, recording and preserving the configuration and parameters (for which there may be multiple instances for specific purposes) is a difficult challenge.

Preservation of these research outputs represents a considerable challenge and is one considered serious enough by JISC and the UK eScience community for them to establish a national Digital Curation Centre [7]. The Centre's work to develop a system to record Representation Information describing data, formats and rendering tools, based on concepts from the OAIS model, may point the way forward. Metadata and in particular, Representation Information is discussed in section 6.

- 4.4 It should also be noted that text outputs can be articles (in various versions), but also working papers, technical reports and laboratory notes. The increasingly data-centric nature of science makes this particular output of special importance and raises issues of granularity. The requirement for the extraction of relevant units of information through technologies such as text mining emphasises the need to include all elements of research in order to answer a query in the most effective way. These various outputs can be referenced within a text document either as linked external files or integrated within so called "compound" or "complex" documents. The relationships between these objects and the ways in which they are used together (e.g. tools, data, parameters and outputs) can be crucial. The outputs produced by the researcher often contain more information than the published version and thus offer the opportunity for re-use or re-purposing, text mining, and other more sophisticated discovery techniques.
- 4.5 There is a question as to what text should be preserved. In certain user communities preprints are submitted to discipline-based repositories such as ArXiv (which covers predominantly physics material) and in some cases institutional repositories as well. Preprints are seen as a way in which the wider community can be alerted to the research and are seen by some as a key method of scholarly communication. Preprints do have the disadvantage of not having been peer-reviewed and are therefore subject to alteration or rejection, potentially damaging individual and institutional reputation. Despite this, it is suggested that non-published material be considered for storage, preservation and ultimately access where possible.
- 4.6 A safer route is to preserve the peer-reviewed version, the postprint. As stated above this could have embedded links to non-textual outputs, enables text mining if in the right format (e.g. XML) and also satisfies recent mandates from funding bodies that the research be made available on open access. This could include conference papers as it could be viewed that these have effectively been discussed within a peer group setting. Postprints often contain more information than their published counterparts. This information is lost if the postprint version is not preserved.

- 4.7 If non-textual outputs are linked to the postprint, then they should be preserved as well, otherwise only a partial view of the research can be discovered. Increasing amounts of data are being produced as a result of e-research and this is being built upon by re-using the data, e.g. in computational science. It may not be feasible to store large datasets within a local repository in which case such data will need to be preserved within trusted data repositories. The inclusion of interim outputs concerned with the process of research rather than the subject of the research itself is not so clear cut. Nevertheless there is likely to be a need, albeit internal to the Higher Education institution in question, for information such as bids, grant letters and collaborative discussions.

## 5. Potential workflow for preserving EVIE research outputs

- 5.1 The following paragraphs should be read in conjunction with the diagram presented in Appendix A. They describe a VRE preservation case study of a VRE hosted at the University of Leeds and supported by a preservation service provided by the British Library. The diagram presents a holistic view of the workflow from the initial generation of the output(s) within the research process supported by the VRE, the export of objects from the EVIE research outputs module, to the discovery of the outputs within the resource discovery portal. While considering some of the specificities of the technology developed within EVIE and the wider infrastructure existing at Leeds (see section 6 for a description of this infrastructure), the case study remains particularly applicable to other HE institutions building VRE infrastructure.
- 5.2 The top-left quadrant of the diagram shows the generation of the initial outputs. It assumes that in the majority of cases this is a text document or that text is the primary format. These are likely to be produced within the document management tool provided within the collaborative environment. A way of managing documents was a requirement from the workpackage 2 user requirements analysis.
- 5.3 In the first instance, this text document will be in its pre-published format, a preprint, and likely to go through a number of versions before submission to the publisher (hence the need for a document management tool). The research might also be disseminated via conference papers, presentations and posters. Eventually the text will reach its final state and become a postprint and if the publisher allows it (as most do), be eligible for deposition and preservation within a local repository. It is quite possible that the postprint will be linked to supplemental data, video and audio and that these will also be deposited in a local repository. In case of the University of Leeds, it is expected that text-based outputs such as postprints will be deposited within the White Rose Institutional Repository and that multimedia will reside in the MIDESS repository. Ideally these would be in digital form to enable delivery from the repository along with the postprint, but this might not always be feasible; in which case the user will have to settle for only discovering metadata pertaining to the object and source the object itself through more 'traditional' channels such as a library. As mentioned above, data might also be submitted to trusted data repositories that are capable of curating this type of output. In either case, there may be a need to preserve the data manipulation software as well as a means of ensuring that it can be used to access the object at some point in the future. Some research results in books being published. These go through similar versioning and collaborative processes to research articles and ideally would also be made available in digital format.

- 5.4 There may be a need to define submission criteria which limit the range of file formats that can be accepted. Some repositories mandate the specific use of file formats such as PDF, HTML and Microsoft Word. Whilst this makes the process of deposition more manageable it can be limiting in that some work is never submitted and preserved, or migration work has to be performed by the submitter. This means that irreversible conversion work is neither controlled nor recorded and there is the danger that the conversion may not necessarily be performed well (i.e. data could be lost). A consensus on this issue has yet to be reached within the HE community with some repositories accepting all formats and others placing strict limits on which formats can be accepted. The DSpace project [24] has considered a compromise whereby formats are grouped with statements as to how well they will be preserved, e.g. Group 1 – guaranteed preservation and representation, Group 2 – preservation will be attempted but not guaranteed, Group 3 – Only bitstream preservation guaranteed.
- 5.5 The EVIE VRE will provide functionality to simplify deposit of research output material, thus enabling the material to enter the preservation workflow. Again, this was a requirement highlighted in the requirements analysis workpackage. The user will enter the basic metadata pertaining to the particular object. Some metadata will be automatically extracted from the object itself, and some may arrive with the object from another source. The object is then uploaded into the EVIE outputs module and depending on format, be deposited within the White Rose or MIDESS repositories. It might be possible to automatically populate the metadata template from a document management system such as Documentum or from data derived from the journal submission process. If the metadata is manually entered, then a quality assurance mechanism needs to be considered to ensure consistency before the data is exposed for searching. This is particularly important in cases where a departmental administrator rather than the author has been tasked with inputting the information. In such instances the quality can be poor. It is suggested that the Leeds University Library continue to act as the focus of the QA process, but do this in a more rigorous manner than simply random sampling; preferably involving the researcher in the sign-off process to ensure compliance with the minimum standard required for RAE 2008. The library should also be responsible for the production of guidelines on the required data standards and organising training for administrators. Batch processing of input should be avoided to prevent workflow problems further down the line. The metadata will also be transferred to the University of Leeds Publication Database, which will be the main source for the discovery interfaces. Metadata for non-digital outputs that cannot be stored in the repository will also appear in the University of Leeds Publication Database (another reason for separating the metadata for resource discovery from the repository database). Such items should be collated by the University of Leeds library so that they can be retrieved from a single source at some point in the future. Outputs may also be sent to external discipline-based or partner repositories as a part of community collaboration. This is already happening at the University of Leeds with the deposition of material into the British Education Index's Education-line service. Metadata issues are discussed in more detail below.
- 5.6 It is possible that there might be material on personal, departmental and project websites that is not deemed suitable for publication. If the sites are deemed important enough then it is likely that they will be harvested as a result of national web archiving initiatives. Where there is no imperative case for preserving certain sites at the national level there may be a valid reason to deposit these locally. Other

forms of material may also be selected for preservation on an individual basis for institutional reasons. An example of this is the personal computer of a leading researcher upon retirement.

- 5.7 The repository used at Leeds is the White Rose Institutional Repository. Whilst it is likely that objects deposited within the White Rose Repository will be available for a number of years, they might not be accessible in the long-term. Policy regarding the repository might change and the maintenance of the objects in terms of migration/emulation might become problematic. The British Library could provide a service to the UK higher education community by storing this valuable material alongside the published version, assuming that it had permission from the various rights holders to do so. Building on its new interoperable platforms it could maintain links with material deposited on trusted repositories. This concept is similar to that employed by the Electronic Theses Online Service (ETHoS) e-theses project [8], whereby theses are imported from UK universities for loading into a British Library repository (currently ePrints). The Preservation Eprints Service (PRESERV) Project [9] and the SHERPA DP [10] project are currently investigating the design of central digital preservation services to institutional repositories, which could range from simple bitstream storage to complex long term preservation support. The outputs of the PRESERV and SHERPA DP are likely to be of great relevance to EVIE. Some form of monitoring procedure for externally preserved objects and means by which continued access can be audited will be required. Work by the Research Libraries Group on the certification of trusted repositories may be relevant here [11].
- 5.8 Work is being done by The British Library on the harvesting of both metadata and content using Web Services and the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) as the means of transmission between the ETHoS university and The British Library. In EVIE this implies the use of a postprints export server linked to the White Rose Repository. At the appropriate point the outputs would be sent to the export server ready for harvesting by British Library systems. The objects would then be transferred to an import server at The British Library where they would be held prior to ingest into the Digital Object Management (DOM) system, the central British Library electronic store. The transfer would be subject to checks and error detection to ensure effective and complete transfer has taken place. The precise mechanism for this process is still under development and is expected to take account of developments coming from the SHERPA DP project and Arts & Humanities Data Service on the possible use of RSS (Rich Site Summary) feeds to identify repository objects and scheduling of the 'pull' mechanism from the export server.
- 5.9 There is a question as to the timing of the export from the repository. It could take place on submission of the item to the White Rose Repository. This would enable immediate discovery of the item from The British Library's resource discovery services and thus make it available globally almost instantly. Alternatively, the item could reside in the White Rose Repository until such time as it was deemed necessary to instigate more comprehensive preservation actions, say after five years. The recommendation is that the item be made available immediately unless an embargo needs to be placed on the object for a set period of time. It is also recommended that this include the complete object, including metadata, to maximise the opportunity for discovery and preservation.

- 5.10 The metadata from the import server would be transmitted to the DOM metadata store, where it will be converted to current DOM standards. Additional, automated metadata extraction may be performed. Persistent identifiers will be assigned to the objects and relationships established between them. Defining the relationship between the published article and the postprint will be key, as both are versions of the research output and the published version may also be stored in the DOM system. The metadata will be surfaced via The British Library's resource discovery services. These will be capable of interoperating with remote portals as well as having their own web-based interface and will therefore be able to expose data within the resource discovery portals of VREs such as EVIE.
- 5.11 The object bitstream will go into the DOM storage subsystem where it is matched with its metadata. DOM is a fully scalable, highly resilient storage system [12]. The DOM Access subsystem will enable retrieval of the objects via British Library delivery mechanisms. At the time of writing, electronic delivery is based on the Library's Secure Electronic Delivery (SED) mechanism, which uses Adobe's PDF secure e-book mechanism to provide controlled access to content that is downloaded to the user's desktop via a hypertext link embedded within a Web page or e-mail. The current SED configuration is quite restrictive. This is due to the type of material currently supplied by The British Library, i.e. published articles. The system could be configured to make it more amenable to less restrictive permissions, such as those espoused by the Creative Commons movement. It is noted that even this may be considered too limiting as far as the University of Leeds is concerned and that a simple header page outlining what can and cannot be done with the document may suffice. The rights are likely to differ based on research sensitivity, domain and type of object. Recent mandates from certain funding bodies demand that outputs derived from research they have funded are made available on open access. It is feasible to assume that outputs relating to the research process might be restricted to University of Leeds staff, and early preprint versions of collaborative documents limited to the research team (e.g. Worldwide Universities Network). The rights management system employed will need to address these various levels of access.
- 5.12 The published article might appear in print or electronic format. Either way it is likely to end up in The British Library collection on shelves or in DOM. If the article appears in one of The British Library's top 20,000 journals then the article will be included in the ETOC database. Although ETOC data can be accessed via a number of channels the main discovery channel for the UK academic community is likely to be ZETOC. This service, hosted at Manchester Computing, provides searching and alerting of the ETOC database free at the point of use for researchers in UK higher and further education. In addition to the web interface, the service provides a level of interoperability based on the Z39.50, openURL, SOAP, SRU/W and Shibboleth standards. This will be developed further as it becomes integrated within e-science and the new JISC e-infrastructure. The ETOC data is also exposed to Internet search engines, currently Google, for those researchers (and there are many of them) that use Google as their main information resource. The Gooraph [13] visualization prototype linked to the Google API developed as part of EVIE workpackage 5 provides an additional interface to aid navigation through records surfaced in the Google result sets.
- 5.13 Although the University of Leeds Publication Database (in some incarnation) is likely to remain the primary local resource for EVIE outputs, possibly supplemented by metadata exposed via the White Rose Repository itself, their exposure through

the various British Library discovery channels mean that the information will be available to researchers across the globe as well as within the EVIE research portal itself. The interjection of The British Library within the preservation workflow effectively ‘closes the loop’ on the research lifecycle that might not otherwise be possible. The involvement of The British Library in the process also raises the question of synchronisation between The British Library and University of Leeds repositories. The Arts & Humanities Data Service is currently looking at how the contents of institutional and preservation repositories might be synchronised. Of particular interest is how the object, or in OAIS parlance its Dissemination Information Package [14], might be requested from The British Library’s preservation store by the originating repository at some point in the future – or whether it should be. A view could be taken that an object preserved by The British Library could be supplied by the Library and that its associated rights information would be sufficient to enable the object to be delivered direct to the user.

## 6. Metadata requirements, standards and issues

### Functional requirements

6.1 An analysis of the various functions of the VRE repository provides some guidance as to the needs for specific types of metadata.

Broad category	Function	Explanation	Metadata
Search	Keyword search	To enable users to perform simple keyword searching based on common description.	Descriptive metadata (eg. Title, Author, Subject, Keywords, etc).
Search	Contextual search	To enable users to discover digital objects based on their context. For example, a user may search for a particular project and then utilise contextual links or structuring to explore the materials available within this grouping.	Contextual metadata that describes the context in which an object was created within or existed within while it was used before it was deposited in the repository.
Search	Search based on origin	To enable users to discover objects based on their provenance.	Provenance metadata describing the origins of the object.
Access	Trust of data	To provide users with some confidence in the quality and accuracy of the information encapsulated in the preserved object.	Provenance metadata describing the origins of the object.
Access	Access control	To allow restrictions to be placed on who can access the object and under what conditions.	Rights management metadata, detailing copyright, owners, rules, etc.
Access	Retrieval	To allow a particular object to be located, retrieved and served to the user.	Unique persistent identifier and associated metadata identifier.
Preservation	Storage and subsequent representation	To allow the structure of a preserved object to be captured, recorded and reproduced.	Structural metadata that captures the relationships between specific files of which a preserved object is composed.
Preservation	Representation	To enable a preserved object to be rendered and used by the user.	Representation Information which describes the file format(s) and mechanisms to understand or render the format(s).

Preservation	Curation and re-use	To enable the user not only to use the object as was originally intended but also to re-purpose or re-apply the use of that object in new ways or with new parameters.	Object specific Representation Information which describes the tools, parameters, data and local context and how the object was used within this context. Could also be broader contextual information.
Preservation	Technology watch	Enables repository manager to monitor for obsolescence and take preservation action as appropriate.	Representation Information, specifically describing the environments within which rendering tools execute, as well as details about the support and maintenance of the tools themselves.
Preservation	Verification of authenticity	Enables verification of the accuracy of bitstream preservation. Allows the effects of preservation action on the preserved object to be recorded and evaluated. May provide feedback on the effectiveness of previous preservation action.	Fixity metadata (eg. checksums) and Provenance metadata (specifically the management history specific to the duration the object has been present in the repository).
Preservation	Significant properties	Enables preservation strategies appropriate to the content to be selected and then applied.	A description of the properties of a digital object which are deemed to be crucial to maintain through preservation action (eg. colour, layout/formatting, interactivity)

**Metadata standards**

6.2 Dublin Core [15] provides a starting point for specifying predominantly descriptive metadata. Basic elements for rights management and contextual metadata are unlikely to be sufficient for the purposes of the preservation of VRE outputs. The inadequacies of Dublin Core for addressing preservation requirements have led to the development of more detailed standards, some of which are considered below. Most elements encountered in Dublin Core are likely to be found in more comprehensive metadata schemas.

6.3 PREservation Metadata: Implementation Strategies (PREMIS)<sup>7</sup> [16] outlines a very detailed set of preservation metadata elements. While PREMIS may adequately support most current and likely future requirements, the challenges of implementing and populating such a large number of metadata elements remain to be answered. Automatic gathering and extraction of metadata is in its infancy. Tools to support the population of PREMIS fields may become available in the future, but for the short term at least, users who submit objects to the repository cannot be expected to enter large amounts of metadata by hand. This is likely to act as a deterrent to submission. It remains to be seen whether it is worthwhile providing a large number of optional fields that in many cases will not be populated. Representation Information is covered by PREMIS in a minimal way, on the assumption that the development of centralised repositories will drive the design of relevant RI metadata elements over the next few years.

6.4 AHDS (SHERPA DP) has published a much reduced metadata set based on PREMIS, containing only 27 metadata “units” [17]. This suggests a more practical limit on the quantity of metadata required may be possible, although this work does not include substantive support for RI.

- 6.5 METS [18] provides an XML based standard for encoding descriptive, administrative, and structural metadata as well as providing a wrapper for encapsulating the content of the digital object with this metadata. It has seen quite wide adoption in part or whole by a range of organisations including the British Library.
- 6.6 MPEG21 provides a framework within which data streams and a variety of XML based metadata can be encapsulated [19]. Much of the impetus of the MPEG21 development is focused on delivery and rights management and industry support will be encouraged by ISO standardisation. Its flexibility lends it some potential for application to preservation functions. This is discussed in more detail by Bekaert, Liu, and Van de Sompel [20]

### **Current metadata standards and practice at the University of Leeds and The British Library**

- 6.6 The White Rose Institutional Repository at the University of Leeds is currently running ePrints software and utilises the standard ePrints metadata schema. This is based on Dublin Core and is augmented with a number of additional fields. These include published status, Refereed Status, Academic Unit (department, unit or school of origin), and an Additional Information field (often containing a copyright statement). Accurately maintaining the Academic Unit has required considerable manual work. Full text searching is provided by ePrints and the most frequent point of access to the repository is Google and Google Scholar.
- 6.7 The University of Leeds Publication Database requires a basic set of fields for predominantly descriptive metadata with some basic support for rule based IPR and privacy control.
- 6.8 The MIDESS Project [21] is currently undergoing a selection process for the software to be used in the MIDESS repository at the University of Leeds. The other two partners are also undergoing a selection process. All partners will use different software to allow[12] comparison and investigation of interoperation. No decisions have so far been made on the choice of metadata schemas.
- 6.9 The British Library is in the process of developing a digital repository in the Digital Object Management Programme (DOM) which will hold the Library's growing digital collections. It is currently reviewing its preservation metadata requirements as it does not consider its current British Library Application Protocol Standard (BLAPS) [22] metadata standard adequate to address the growing digital access and preservation challenges. The DOM programme has adopted METS as a metadata framework. It is developing an object model to describe structural relationships between digital objects [12]. The British Library is involved in detailed metadata discussions as part of its liaison with the Legal Deposit.

### **Addressing metadata requirements**

- 6.10 Descriptive metadata enabling keyword based searching is reasonably well standardised and will be found in the shape of Dublin core elements in the default metadata schemas provided in repository software like GNU ePrints [23] and DSpace [24], and used in existing repositories at the University of Leeds.

- 6.11 The particularly complex areas of Representation Information for long term preservation and Context Information to target curation purposes are also at an early stage of experimental development. Support for these crucial areas is likely to be developed over the next few years as the activities of projects, institutions and organisations involved in digital preservation and curation work progress. The developments of the Digital Curation Centre on Representation Networks [25] and curation, the PRESERV Project’s work on preservation services, and the National Archives’ [26] and Harvard’s work on automatic digital object identification and validation [27], show potential ways forward
- 6.12 Standards for rights information are starting to emerge. A Rights field already exists within Dublin Core, but much work is being done building on the INDECS (Interoperability of Data for Electronic Commerce Systems) project [28]. Rights information is vital to enable delivery of objects to the appropriate user and to prevent abuse. In order to develop integrated, seamless delivery, the metadata will need to be machine interpretable. The importance of this area for British Library services has been recognised and the Library is currently working with third parties to inform its implementation of rights management within the Digital Object Management programme.

**Approaches to metadata capture and organisation**

- 6.13 Selecting the appropriate level of detail in the metadata schema used is likely to be a compromise between adequately fulfilling the functional aims of the repository and the practicalities of populating the metadata fields. As techniques for gathering and recording metadata improve, the level of detail that can be stored and utilised can increase. Various approaches are listed and discussed below

Method of gathering metadata	Description	Issues
Manual entry	This typically involves the user who submits an object to an archive, entering details into a number of fields by hand.	Manual entry must be kept to a minimum to avoid putting off depositors. Fields populated in this way must include the key information that can only be sourced from the likely contributors (often the authors in the case of a VRE). Guidance must be given as to how the fields can be effectively completed.
Repeated related entry / wizards	Where a user submits a number of objects it is likely that a number of metadata fields are likely to be similar for each object. Hence the interface provides an option to auto-complete fields as per previous entry.	It is likely that while some fields will be completed with the same entries for a number of objects submitted by the same user (eg. author field), others will still have to be entered by hand (eg. title field). The technique offers relatively powerful returns from simple interface development.
Automatic extraction	Automated tools executed on ingest extract metadata from the submitted object.	This technique can be used in a number of areas. A tool could extract Title and Author metadata where this is present in the deposited object (eg. Word or TIFF files may contain this). This could then be presented to the user to be amended or confirmed. The National Archives’ DROID tool offers a different example, where an object’s file format is identified automatically. This technique is particularly useful for extracting technical metadata that the user may not have the knowledge or skill level to ascertain for themselves.
Packaged metadata	Some objects deposited in the repository may already have had metadata created for them. Perhaps the object had already been submitted to a subject specific	Packaged metadata will not be immediately useful for searching or performing other functions, but it is possible that support in existing software is extended or even that extraction tools of the future will do a better job of utilising existing metadata. While

	archive. While it may be possible to automatically extract and re-use metadata, whatever form it is in, a very practical and simple technique is simply to package the metadata and attach it to the deposited object. It may prove useful to return to (perhaps utilising automatic extraction tools) at some point in the future.	depending on this technique without recording metadata more thoroughly would be somewhat dangerous, it is simple enough to implement just in case the packaged metadata proves to be useful later. If the metadata obtained with an ingested object cannot be entirely recorded within the repositories metadata schema, this strategy may be an appropriate compromise.
Indirection approach	While some metadata fields will be completed with details specific to each object in the repository, many will have identical entries. In some cases it may therefore make sense to store this metadata separately and reference it from each object as appropriate. For example, many objects could have the same author information and simply point to this metadata using unique identifier technology.	Representation Information is a key area where this technique is likely to be useful. Repeating metadata of this kind for each object in an archive will be impractical as inevitably it will evolve over time. Context information is another contender for this approach which would significantly reduce the effort required to create and manage the metadata while improving the accuracy and ability to search it. For authors moving between institutions a rigorous naming scheme is needed.
Externally referenced	Some metadata may be provided externally that is considered both accurate and trustworthy. Where appropriate this metadata could simply be referenced rather than duplicated.	Taking the “Granular” approach further, metadata could not only be referenced from a number of objects in the repository but it could even be managed externally from the archive. For example, Representation Information describing the PDF format and how it can be rendered is likely to be of great use to many repositories and may be provided as an external service. The latest version of PRONOM [29] utilises unique identifiers as a first step to realising this technique.

## 7. Strategy for developing metadata and long term preservation functions in EVIE

- 7.1 EVIE recommends a practical approach to address the functional needs for recording metadata in the most effective way possible in the short term. It will outline a strategy for enhancing support for these functional needs as standards, technology and preservation services become available over the longer term. Awareness of likely developments will ensure that a flexible approach can be utilised that will avoid problems in the medium and longer term.
- 7.2 Support for simple keyword searching can be fulfilled by the standard descriptive elements provided in the existing metadata schema of the Leeds repositories. The existing ePrints software also provides support for full text searching and this is exploited by many users via Google.
- 7.3 As described above, contextual searching and curation functions are yet to be supported in depth by existing metadata schemas. The technology to implement these functions is also in its infancy. Basic contextual and provenance metadata is captured using the University of Leeds’ “academic unit” field. In the short term, this could be extended to provide further levels of detail (eg. Unit, Group, Project), if a cost effective strategy for maintaining this metadata can be established. This would support contextual browsing across related VRE outputs following initial discovery. This will become increasingly important as a greater range of VRE outputs are deposited rather than the current focus on pre-prints.

- 7.4 Looking to the medium and longer term it is likely that more comprehensive support for contextual metadata will be provided and this will ideally be adopted at an appropriate time. Subject specific metadata and sophisticated techniques for describing contextual and structural metadata are likely to facilitate more effective methods of resource discovery in the future. Providing a stronger solution to recording structural metadata is non-trivial and will hopefully become available from developments to the ePrints software used in the White Rose Repository.
- 7.5 Facilities and support for recording and utilising comprehensive Representation Information (RI) are likely to be beyond the realistic resources of an institutional VRE at the current time. However, as described above, currently there is work developing centralised RI repositories and these are beginning to be populated. A sensible approach is therefore to reference RI externally where suitable trusted metadata becomes available and contribute metadata content where local expertise and effort is available. Implementing an automatic file format identification tool like DROID or JHOVE and adding metadata fields for storing the key results will be a useful starting point. PREMIS suggests the use of fields to capture an object's format and version, while referencing external RI in fields describing the format repository and the method of identifying the format in that repository. Effective long term digital preservation actions are unlikely to be employed until cost effective preservation services are provided externally. Characterising digital objects will provide a useful foundation for the preservation action which will need to be taken in the future.
- 7.6 In the short term, simple extensions to the existing metadata schema used at the University of Leeds in the White Rose Repository will provide adequate support for the functional requirements described above, given the limitations of current technology and preservation services. In the medium to longer term, technology, services, and metadata to support these functions will become available. Standards for preservation metadata schemas like PREMIS will be more established and will have been more practically tested. Wider uptake will help to inform when these possible standards and services might be adopted.

## **8. Recommendations for digital preservation within EVIE and other VREs**

Although the workflow described in this document focuses on the University of Leeds infrastructure, the overall concept could apply to any research institution implementing a VRE and maintaining a local repository. Many of the issues associated with the local preservation of objects over the long-term could be avoided if the EVIE workflow were adopted and The British Library used as the preservation store and the local repository focusing on local administration needs and resource discovery.

- 8.1 A policy document should be produced stipulating what research outputs are to be deposited in local and subject repositories. This should explain the benefits to the organisation and its staff of deposition in terms of RAE2008 requirements, compliance with funding mandates and visibility within the emerging global discovery infrastructure.

This recommendation is a fundamental component of the model described in this document. There is unlikely to be a major increase in deposited items in a local repository unless potential depositors are aware as to why it exists and the role it plays within the organisation. The importance of the repository in supporting the institution with regard to its credibility, funding and visibility (and similarly its researchers) has to be put across. It should be explained that an article is often only visible to those who subscribe to the journal in which it is published; even open access articles can lose visibility unless they can be discovered via popular interfaces such as Google or Web of Knowledge or from search technologies embedded within the user's desktop environment.

Recent mandates from the Wellcome Trust and other funding bodies [30] demand that research is made available on open access. As a result, research is likely to find its way onto open local repositories or open subject repositories such as PubMedCentral. This may well force the issue and make certain that items are deposited, but provision still needs to be made to ensure that research that falls outside such mandates is also included. It needs to be made clear that items not in the repository will not be harvested and therefore not be disseminated through major discovery channels.

- 8.2 The range of material deposited in repositories should be expanded to include preprints, postprints and associated supplementary material. A mechanism needs to be put in place to enable the institution to monitor research outputs and submissions. Justification must be made for non-deposition (e.g. copyright restrictions or below quality threshold).

The impact of this recommendation is dependent on the purpose of the deposition and how the data is to be used. Early versions of preprints and material derived from the research process itself might be restricted to internal use. An institutional repository could therefore be seen as simply a shared server within the organisation with little requirement for long-term preservation. Such restrictions would suggest the need for access control to ensure that only certain groups of staff are able to access the material. There is an issue regarding the dissemination of non-peer-reviewed material and the possible impact on an institution if the work is discredited during the review process. The counter-argument to this is that by posting preprints onto open repositories the community is alerted to the fact that the research exists. This is an argument embraced by the physics community and is one that should be given due consideration wherever possible. In this instance the status of the document needs to be made very clear. Postprints, being peer-reviewed, are the research outputs most likely to be found in repositories and are increasingly considered to be more valuable than the published version due to the extra information (e.g. full datasets) that is sometimes included. Postprints on open access could increase demand for, and visibility of, the research.

More effort will be required to deposit this material. Mandatory deposition of research outputs will obviously increase the workload of the researchers and their assistants/departmental administrators. Training on the process will be required, as will on-going support. The institutional library would seem to be ideal place to focus this activity. A single focus would help ensure consistency. Mandatory deposition also suggests that the library would need to be aware of the outputs that are being produced. This is unlikely to be a trivial task, requiring the monitoring of research projects and their expected deliverables. Whilst it might be worth exploring the use of a set of central servers and directories where outputs are stored by default, as used by some organisations, and the use of wikis to ensure compliance and aid monitoring, realistically it is a user education issue as per 8.1. The result of this is likely to be an increased workload on the local library and the possible requirement for more staff.

- 8.3 A list of preferred formats for repository objects should be issued, with explanation as to the way in which various formats will be preserved. The list of formats should be regularly reviewed. The institution (possibly the library) should also establish a mechanism for the physical conversion of non-preferred formats to preferred ones to help provide consistency across the repository and encourage researchers to deposit.

Most repositories appear to be settling on PDF, MS Word and HTML versions of documents. Some are looking at XML as a possible longer-term alternative. The British Library and Library of Congress are supporting the NLM-DTD. Conversion to this DTD is a non-trivial process, but if the preservation of the object is to reside with The British Library then the NLM-DTD would be the preferred option. The British Library could offer the conversion to NLM-DTD as a service to institutions. It could also undertake Word to PDF conversion, assuming that both the XML and PDF versions would be preserved. This would have less of an impact on the institution as it need only be concerned with ensuring that the documents conformed to a standard acceptable to The British Library. It is unlikely that any such conversion would be carried out by the user and the institution's library might be required to carry out the intermediate steps. The 'annual review' would therefore need to take into account The British Library's latest formats. The above focuses on textual objects, but the principle would apply to other objects as well.

- 8.4 Repository items should be exposed to internal and external discovery channels, including those of The British Library, on submission. This should be the norm unless the item has been specifically embargoed.

The default would be to expose repository objects to the institutional portal/VRE. It is assumed that this matter would have been agreed and potential solution addressed before embarking on a scale-up of the local repository. The reliance by researchers on Google indicates that the data could be harvested by Google and made available via its search interface. But such harvesting might not be consistent in the long-term and the required dissemination of the institution's research might not happen. A similar argument applies to other initiatives such as Elsevier's Scirus [31]. A more consistent approach might be to allow The British Library's copy of the object to be harvested by internet engines, thus ensuring (as far as is possible) that the object will be exposed via these channels. By default, The British Library's copy of the object would be surfaced via The British Library's own search interfaces and through its relationships with other organisations such as OCLC (e.g. Find in a Library). Such an approach reinforces and supports the visibility of research as outlined in 8.1.

- 8.5 A mechanism should be established to ensure that non-digital research material is catalogued and curated. The metadata should be discoverable alongside that of digital objects.

This is likely to have an impact on staff resources within the institutional library. By definition, these objects are not likely to be amenable to the automated metadata generation or storage mechanisms described in this document. The cataloguing of these items will be manual, as will their long-term management. There might also be physical storage issues, not only in terms of 'shelf' space but also the conditions under which tape etc needs to be kept. It is nevertheless vital that such research information is not lost. As users become accustomed to accessing digital audio and video, so it will raise expectations that similar information in non-digital format will also be available – if not quite so readily. Metadata will have to be generated in order to expose these objects through the

discovery channels in 8.4. The task may not be too onerous; much depends on the amount of material that will actually be produced this way (e.g. on cassette tape as opposed to recording directly onto hard disk). The first task will be to audit the current set of outputs and estimate the likelihood of future non-digital outputs. The costs of maintenance might indicate that it is more cost effective to invest in digital technology at the outset (e.g. immediate download from a hard disk camcorder rather than shelving a DV tape).

- 8.6 The metadata relating to repository objects should be quality assured by the institution's library. It should take responsibility for the training of those tasked with inputting metadata (e.g. departmental administrators) and produce guidelines on the minimum standards required for assessment (e.g. RAE2008). A sign-off process involving the researcher should be implemented.

The institution's library should assess the requirements of the RAE with the appropriate research unit and decide on the minimum acceptable standards to meet the RAE. A series of training sessions is required to ensure that all those inputting metadata (e.g. departmental administrators) are trained. It might be necessary to enforce a procedure by which only those accredited by the library can input into the system. The library should institute a regular check of the metadata (random sampling if the numbers are large) to ensure consistency. To minimise the effort later on, it is important that it is right first time. The process must therefore include a sign-off mechanism by the researcher before data is loaded on to the system. This could be a massive educational task – but again it must be made clear to researchers that their visibility and reputation is on the line if incorrect information is widely exposed; so it is in their own interests to get it right.

In order to get buy-in, the process should be as simple as possible, preferably using an 'I Accept' button on the keyed metadata, with an option to edit if necessary. Such a mechanism would enable the library to monitor those that have not been signed-off and for which 'user education' is required. Although the ideal is to make it easy to submit, a balance needs to be struck between ease of submission and the value once deposited. At the very least, descriptive metadata is required. Other data fields may have to await the outcome of the various developments outlined in 8.13.

- 8.7 Wherever possible common interoperability standards (e.g. OAI-PMH, Web Services) should be adopted to ensure compatibility with British Library systems.

This is evident if the preservation model presented in this report is to be adopted. It is unlikely that any implementation will deviate from JISC e-Framework standards; the use of such standards will ensure interoperability with The British Library (the Library is a member of the Common Information Environment group). The standards are likely to become clearer as a result of work carried out under the 2006 JISC repositories and infrastructure capital programmes, but would be expected to embrace those specifically mentioned in the metadata sections of this document such as RSS, Dublin Core and METS.

- 8.8 Selected repository objects should be exported to The British Library shortly after submission to maximise global discovery and the opportunity for long-term preservation. In the first instance, it is suggested that only peer-reviewed objects be submitted to the preservation workflow.

The benefits of exposing through multiple discovery channels is covered in 8.4. The earlier the object is sent to The British Library the less effort is required by the institution to preserve the object. Until the quality and political issues are sorted, it is safest to go with peer-reviewed material for preserved objects. The numbers are likely to be more manageable.

- 8.9 A procedure should be put in place for the monitoring of objects that have been preserved in systems external to the institution's local repositories. This should include a means by which continued access to such objects can be audited.

The workflow proposed in this document implies that some sort of auditing process is required in order that the institution can be assured that their material is being handled effectively. This will be particularly important until a base of trust has been established. In the first instance, this might be a manual consistency check carried out by library staff on The British Library's system (using an account specifically to allow depositing institutions to monitor their own material, which in some cases may be embargoed for a period of time). Rights management is therefore essential to ensure that the appropriate institution and individuals within the institution have immediate and continuing access to their material. In the future, an automated process should be implemented matching rights metadata on The British Library object with identification of the requester.

- 8.10 The rights and restrictions expected to be assigned to the objects populating local repositories should be identified. These should be used to establish a rights policy for research outputs sufficient to maximise access and discovery but prevent unwanted commercial exploitation and plagiarism. The policy will need to address recent mandates from funding bodies, internal confidentiality, collaborative requirements/restrictions and delivery of the object from repositories external to the institution such as The British Library (e.g. using Creative Commons and/or Secure Electronic Delivery).

This is likely to be difficult but vital. There is not only the question of identifying specific staff currently working at the institution and with the right to access their own documents and those of their research group, but also those that have left the institution but still have the right of access. The rights have to address the rights of the authors but also the publishers, including any embargo period relating to dissemination, whether that be by secure delivery, Creative Commons or open access. Such access is likely to change over time. Different versions of an object could have different rights (e.g. to avoid pre-publication plagiarism or pending formal investigation). Even using Creative Commons decisions need to be made regarding what level of permissions are to be granted (e.g. third party mining might be acceptable to identify 'nuggets' of information, but not wholesale copying for input into a published book). Rights become even more complex as research becomes more collaborative. Collaborating institutions are likely to have different views on the use of their material. All of this needs to be mapped onto the associated delivery mechanism of The British Library's preservation system. A review of current research outputs needs to consider the rights that ideally would be allocated to each one and for research Deans to suggest how such decisions should be made and the criteria for assignment.

- 8.11 The feasibility of implementing simple extensions to repository schema to provide support for preservation services should be assessed.

This is partly related to 8.10. It is suggested that institutions requiring long-term preservation services from The British Library take the lead from the Library in terms of the required metadata. Although technical metadata could be automatically generated by The British Library, it is unlikely that it would be able to assign all the necessary metadata without input from the creating organisation, e.g. the level of granularity associated with an 'academic unit'. The institution should also consider adopting a file format identification and validation tool and extend the metadata schema to hold this information. Ideally this should await the outcome of investigations in the PRESERV and SHERPA DP projects before being progressed. The referencing of data likely to be reused (e.g. author information) should be stored separately from the object to aid in the long-term maintenance of the data.

#### 8.12 Monitor developments in the following areas, and adopt as appropriate:

Metadata extraction tools	Tools such as the PREservation Metadata Input Tool (PREMINT) [32] may provide useful mechanisms to populate comprehensive preservation metadata schemas.
Preservation Metadata Schemas	The availability of supporting tools and practical testing at other institutions of standards such as PREMIS and MPEG21 will inform possible adoption
Repository software	Developments of the software currently used in a number of UK repositories (eg. as part of the PRESERV and SHERPA DP Projects) are likely to provide the simplest routes to developing support for preservation
Preservation services	Externally provided services (as being considered by the PRESERV Project) may provide cost effective preservation solutions
RI repositories	Referencing external Representation Information will provide a cost effective way of ensuring archived objects can continue to be rendered over time
Cost effective methods of preservation action (eg. Migration on Request [33])	On demand tools and techniques may provide economical alternatives to external preservation services in some cases

The above demands an ongoing technical watch on developments. The institution therefore needs some skills to understand the developments and be able to act on them. Again The British Library could help here and provide advice on the tools and standards that should be adopted and that comply with its own policies. Nevertheless, some training

could be required in the local library in order to get the necessary basic skills. The level required would depend on whether the institution was handling its own preservation or whether it was outsourced to a trusted subject repository or The British Library.

## 9. References

- [1] User Requirements Analysis Report,  
[http://www.leeds.ac.uk/evie/workpackages/wp2/evieWP2\\_UserRequirementsAnalysis\\_v1\\_0.pdf](http://www.leeds.ac.uk/evie/workpackages/wp2/evieWP2_UserRequirementsAnalysis_v1_0.pdf)
- [2] Systems Integrations Requirements Analysis,  
[http://www.leeds.ac.uk/evie/workpackages/wp3/EVIEWP3\\_SysRequirements\\_v0\\_4dms.pdf](http://www.leeds.ac.uk/evie/workpackages/wp3/EVIEWP3_SysRequirements_v0_4dms.pdf)
- [3] Excuse Me... Some Digital Preservation Fallacies?, Rusbridge C, Issue 46 February 2006, <http://www.ariadne.ac.uk/issue46/rusbridge/>
- [4] A blueprint for Representation Information in the OAIS model, Holdsworth D, Sergeant D, Eighth NASA Goddard Conference on Mass Storage Systems and Technologies, <http://esdis-it.gsfc.nasa.gov/MSST/conf2000/PAPERS/D02PA.PDF>
- [5] PLANETS, <http://www.planets-project.eu/>
- [6] CASPAR, <http://www.casparpreserves.eu/>
- [7] Digital Curation Centre (UK), <http://www.dcc.ac.uk>
- [8] EThOS e-theses project, <http://www.ethos.ac.uk/>
- [9] PRESERV Project, <http://preserv.eprints.org/>
- [10] SHERPA DP Project, <http://ahds.ac.uk/sherpa-dp/>
- [11] Trusted Digital Repositories: Attributes and Responsibilities, Beagrie, Doerr, Hedstrom et al, <http://www.rlg.org/legacy/longterm/repositories.pdf>
- [12] Design for the Long Term: Authenticity and Object Representation, Farquhar A et al: <http://www.bl.uk/about/policies/dom/pdf/archiving2005l.pdf>
- [13] GooRaph: Document Visualization of Search Results, Yiqing Xu, Gemma Kitchen, Derek Sergeant
- [14] Open Archival Information System, ISO, <http://nost.gsfc.nasa.gov/isoas/>
- [15] Dublin Core, <http://dublincore.org/>
- [16] PREMIS, <http://www.oclc.org/research/projects/pmwg/>
- [17] Proposal for a minimum preservation metadata element set based upon the PREMIS data dictionary, Knight G, [http://ahds.ac.uk/about/projects/hybrid-archives/wp44\\_preservation\\_metadata.pdf](http://ahds.ac.uk/about/projects/hybrid-archives/wp44_preservation_metadata.pdf)
- [18] METS, <http://www.loc.gov/standards/mets/>
- [19] MPEG 21 overview <http://en.wikipedia.org/wiki/MPEG-21>
- [20] Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library, Bekaert J, Hochstenbach P, Van de Sompel H, <http://www.dlib.org/dlib/november03/bekaert/11bekaert.html>
- [21] MIDESS Project, <http://www.leeds.ac.uk/library/midess/>
- [22] Dublin Core Application Profiles at the British Library by Robina Clayphan and Bill Oldroyd,  
<http://dc2005.uc3m.es/program/presentations/Tuesday%2013.%2012.00h%20-%20ROBINA%20CLAYPHAN%20-%20BILL%20OLDROYD.ppt>
- [23] EPrints, <http://www.eprints.org/>
- [24] DSpace, <http://www.dspace.org/>
- [25] Digital Curation Centre Representation Registry, <http://dev.dcc.ac.uk/dccrrt/>
- [26] DROID, <http://www.nationalarchives.gov.uk/aboutapps/pronom/droid.htm>
- [27] JHOVE, <http://hul.harvard.edu/jhove/index.html>

- [28] INDECS, <http://www.indecs.org/>
- [29] PRONOM, <http://www.nationalarchives.gov.uk/aboutapps/pronom/puid.htm>
- [30] Recent mandates from the Wellcome Trust and other funding bodies,  
<http://www.rcuk.ac.uk/access/2006statement.pdf>,  
[http://www.wellcome.ac.uk/doc\\_wtd002766.html](http://www.wellcome.ac.uk/doc_wtd002766.html)
- [31] Scirus, <http://www.scirus.com/>
- [32] PREMINT, <http://metadata.net/panic/Results/questionnaire/tool.htm>
- [33] Migration on Request, <http://www.si.umich.edu/CAMILEON/reports/mor/index.html>