Strategic Content Alliance Standards Guide

DRAFT

November 2008
Contents

Scope of the Standards Guide ................................................................. 4
Background ............................................................................................ 4
Why use standards? .............................................................................. 5
Context of use........................................................................................ 5
Glossary ............................................................................................... 6

Access management ........................................................................... 7
What does Identity and Access Management do? .............................. 7
LDAP .................................................................................................... 7
Shibboleth and SAML Standards .......................................................... 7
UK Access Management Federation .................................................... 8
Eduperson Standard ........................................................................... 8
XACML ............................................................................................... 9
References and further reading .......................................................... 9

Metadata ........................................................................................... 11

Technical and Administrative Metadata ............................................. 14
Image metadata .................................................................................... 14
Multimedia metadata .......................................................................... 15
Structural metadata ............................................................................ 15
Bibliographic, descriptive or delivery metadata ................................. 16
Preservation metadata ........................................................................ 18
References and Further Reading (on Metadata) ................................. 19

Digitisation ......................................................................................... 21
What do we mean by digitisation? ...................................................... 21
Why digitise? ..................................................................................... 21
Preparation for digitisation ................................................................. 21
Capture and conversion ..................................................................... 22
Geographic Information Standards (GIS) .......................................... 25
Mark-up ............................................................................................ 25
Page presentation mark-up ................................................................. 27
References and Further Reading ........................................................ 28

Digital Curation and Preservation ....................................................... 29
Preservation Related Standards ........................................................ 30
Persistent Identifiers .......................................................................... 31
Multiple systems of identifiers ........................................................ 33
Questions and issues ........................................................................ 34
References and Further Reading ........................................................ 34

Interoperability .................................................................................. 35
Service registries ................................................................................ 37
Web services ...................................................................................... 39
References and Further Reading ........................................................ 40

Resource Discovery ............................................................................ 41
Distributed searching .......................................................................... 42
Digital repositories ............................................................................ 43
Semantic Web Services ...................................................................... 43
Emergent Standards .......................................................................... 44
Common Vocabularies ..................................................................... 44
Portals and portlets .......................................................................... 45
Presentation layers ........................................................................... 45
Location and delivery ....................................................................... 46
References and Further Reading ........................................................ 46
Conclusions for Use of Standards ................................................................. 47
Recommendations for Use of Standards .................................................. 47

Glossary of terms ..................................................................................... 48
Scope of the Standards Guide

This document has been commissioned through the work of the Strategic Content Alliance (SCA) to synthesise current information about standards and their use in five categories: Access Management, Digitisation, Digital Curation and Preservation, Interoperability, and Resource Discovery. One of the key areas of the UK content Framework is the use of standards in a networked environment, ranging from capture, conversion and mark-up used in the digitisation of analogue materials through to the curation of digital objects for long term storage. It is intended that this will be a first step towards a comprehensive standards policy covering all areas of JISC work including e-learning and e-research. Standards specific to these areas are not currently covered in this document.

The JISC Standards Catalogue\(^1\) provides entries on key standards, including a risk assessment for their use. The catalogue collates information from experts from JISC services and other interested and expert users to provide new editions created over time in an editable wiki. This Standards Guide is intended to enrich the information that can be found in the JISC Standards Catalogue. The catalogue allows quick reference access to information on standards, and is intended for use by programme managers and technical staff. This document is aimed at staff managing development projects, and library managers, and provides an easy to read summary of the main issues surrounding the use of certain standards, and the pros and cons for their adoption. There should also be recognition that information about standards and how they are being used is continuously developing.

It is also intended to put the standards themselves into an easily understandable context. Therefore, this document also shows, by case study example where possible, how certain standards have been applied by institutions in receipt of JISC project funding. Through context of use we also examine concepts such as usefulness, fitness for purpose and maturity of the standards.

The categories chosen are aligned to and complement the work of the JISC e-Content\(^2\) and Information Environment\(^3\) areas. Standards are not exclusive to any one category and many will fit into more than one according to their role. For example, the OpenURL standard is a means of capturing information and also a means for communicating this information between systems.

Background

Standards arise and are developed for many different reasons. The definition of a standard used by the British Standards Institution (BSI) is as follows: “A standard is a published specification that establishes a common language, and contains a technical specification or other precise criteria and is designed to be used consistently, as a rule, a guideline, or a definition. [Standards] help to make life simpler, and increase the reliability and the effectiveness of many goods and services we use.”

As the MINERVA Technical Guidelines\(^4\) state, the appropriate use of standards can deliver the consistency that makes interoperability possible. A key part of what the JISC as an organisation aims to achieve is interoperability, which at its core results from a high level of consistency across the digital resources made available by multiple providers. If tools and services operating across multiple resources need to handle only a limited number of clearly specified formats, interfaces and protocols then this should facilitate rapid development, and should cut costs.

The eReSS team identified\(^5\) that the barriers to the implementation of standards are varied, but if they can be identified then duplication of bad experiences may be prevented. The process through which standards are developed means that they capture good practice based on past experience and enforce rigour in current practice. In the same report from the eReSS project the concept of a standards lifecycle and usage is raised. The lifecycle traces the life of a standard from conception through to ratification and in some cases obsolescence. Both the barriers and the stage of the standard’s lifecycle will influence decisions about when and where to adopt a particular standard.

---

\(^1\) [http://standards-catalogue.ukoln.ac.uk/index/JISC_Standards_Catalogue](http://standards-catalogue.ukoln.ac.uk/index/JISC_Standards_Catalogue)

\(^2\) [http://www.jisc.ac.uk/aboutus/committees/sub_committees/cs.aspx](http://www.jisc.ac.uk/aboutus/committees/sub_committees/cs.aspx)

\(^3\) [http://www.jisc.ac.uk/whatwe/do/themes/information_environment.aspx](http://www.jisc.ac.uk/whatwe/do/themes/information_environment.aspx)


\(^5\) eReSS Study, Initial phase report: Standards and technologies in the VRE domain. September 2006
Most standards used in development work are ‘de facto’ standards, technical approaches that have become widely used through practice and word of mouth. Other standards are ‘de jure’, formally recognized by an organisation responsible for setting and disseminating standards and developed to meet a set of requirements (e.g. the TCP/IP\(^5\) set of protocols).

Some standards are closed, managed by their creators but not available for the community to change (e.g. Adobe’s Portable Document Format (PDF) standard); others are open standards, developed in a way such that a broader community can contribute and ensure that the final outcomes are driven by their requirements as well as those of the standards provider. Within this category are standards which have open access, to both the standard and documentation produced during its development, and open use, where implementing the standard is managed through licensing.

In a nutshell, standards are essential in the production of electronic resources because they facilitate data interchange and representation and management of information. That appropriate standards should be used in research and development projects is almost unarguable; which standards should be used is another question. As far as open standards are concerned, there is even more debate\(^7\). Further, with the recent popularity in reuse and repurposing of digital resources, it is likely that a lightweight approach to standards, which is good enough, rather than perfect, may facilitate reuse as well as development of “mash-ups”.

Alternatively, good practice guidelines, as opposed to “heavyweight” standards, may be useful if they provide genuine guidance without excessive compulsion\(^6\); the risk is that the guidelines may be either so precise as to be restrictive, or so general as to fail to provide useful guidance. They must be capable of application in the local context. They also have to be credible, preferably with clear evidence to support them. They should not be used to the exclusion of other techniques, in particular testing with users.

Why use standards?
The main benefits for using standards are:

- To avoid duplication of effort;
- To build upon the work and best practices of other organisations;
- To facilitate exchange of information between organisations using the same standards (interoperability); in particular for data and metadata standards;
- To demonstrate quality assurance through use of mature standards;
- To potentially reduce costs.

Context of use
When any new technical solution is implemented it is always valuable to understand how others have managed this already. This is particularly the case with standards and specifications: an incorrect implementation can prevent the interoperability that the standard is intended to provide. Quite often it will be the implementation of a related technology that is the focus of project work. For example, in the implementation of a Shibboleth instance, being able to get assistance from those who have already addressed many of the issues involved can save time and effort\(^9\).

---

\(^5\) See http://www.ietf.org/rfc/rfc2151.txt
\(^7\) e.g. http://www.minervaeurope.org/publications/MINERVA-Technical-Guidelines-Version-1.2.pdf
\(^8\) See http://iamsecd.ncl.ac.uk/ and reports from Shibboleth early adopter projects http://www.jisc.ac.uk/whatwedo/programmes/programme_cminfrastructure.aspx
Glossary
A glossary of terms has been prepared supplementary to this guide. This simplifies the definitions of terms used in the document, and defines some terms that are not defined in the text of the guide for reasons of ease of reading of the guide.

Note: Wikipedia
We have avoided referencing Wikipedia as it is a non-standard source. However, the reader is encouraged to carry out their own further reading, and this may include Wikipedia pages. For some emergent technologies Wikipedia may be the most up-to-date resource available.
Access management

What does Identity and Access Management do?
Identity and access management (IAM) ensures that the right people access the right services. In the past, this was implemented system by system with duplicate identity data distributed across institutions. When another service was added, so was the identity infrastructure to go with it. The distributed security issues associated with these duplicate identity stores have become unmanageable.

The solution is to use the same identity information service for all applications. This:

- simplifies by using one IAM infrastructure over and over;
- secures by consolidating identity infrastructures from many to one and reducing the security issues from overwhelming to manageable.

The first step in building an IAM infrastructure is to review the data about people, decide what’s relevant from the source systems, and consolidate and update the information into one identity entry for each person in the community.

Federation dominates access management at present. Federation refers to the establishment of trust agreements and the identification of users or attributes across security and policy domains to enable more seamless interactions. Where web services aim to enable integration through loose coupling at the application and messaging layer, federation does so at the identity management layer. In so doing, federation also isolates each domain, and keeps private the details of the others’ authentication and authorisation infrastructure.

LDAP
The Lightweight Directory Access Protocol (LDAP) is a technology that enables distributed access to remote directories of information. Developed initially to provide access to telephone directory-type information (name, address, telephone number, etc.), LDAP is now more often used to provide access to a diverse range of standalone databases holding information about people, organisations, services, Internet resources, etc. LDAP is often used by other services for authentication, despite the security problems this causes.

Users can be identified through LDAP. However, there has been confusion surrounding single sign-on and the use of LDAP look-up. Some institutions have claimed they have single sign-on by having a single LDAP database which a number of services access. However, this is not true SSO as the user will still be challenged individually by each service.

Shibboleth and SAML Standards
The standards-based Shibboleth software has been developed by the Internet2 community in the United States. Shibboleth defines a common framework for access management that is being adopted by education and commercial sectors across the world. Like many open source software packages, Shibboleth uses, and informs the development of, standards. The most relevant ones are Security Assertion Markup Language (SAML), which is used for the passing of information about access by Shibboleth; and eduPerson, an LDAP object class for describing individuals in Higher Education.

Shibboleth does not carry out authentication itself. Instead, Shibboleth defines a set of protocols for the secure passing of identity information between institutions and service providers. It relies on the institution to

10 http://www.tech-faq.com/ldap-security.shtml
11 http://www.jisc.ac.uk/uploaded_documents/CMSS-Gilmore.pdf
12 http://shibboleth.internet2.edu/
establish identity, and on the service provider to confirm access rights, given information about institutional affiliation. It is written in SAML (Security Assertion Mark-up Language), an international standard developed by the OASIS Security Services Technical Committee 13.

SAML is an XML-based framework for communicating user authentication, entitlement, and attribute information. As its name suggests, SAML allows entities to make assertions regarding the identity, attributes, and entitlements of a subject (an entity that is often a human user) to other entities, such as a partner institution.

In April 2008 Internet2 announced a major release of Shibboleth, which added new encryption features from the SAML 2.0 standard, and improved ease of implementation and management of anonymous identifiers. Shibboleth 2.0 provides an improved method for usage logging at the home institution to better track abuse or inappropriate use of the system. To protect user privacy the identity provider can assign a persistent unique identifier to a specific user. This allows service providers to tailor services to the needs of that user without knowing the specific identity. For instance, a student researching articles in an online journal could save his searches using the anonymous identifier and build on the research over time. This is transparent to the user, who does not have to know the identifier.

From an operational perspective, the new version of Shibboleth makes it easier for information technology staff at the identity and service provider institutions to install, operate and manage the software. Shibboleth 2.0 allows institutions to use their legacy directory schema by translating the data into the federation-specific attributes as needed in real time, greatly decreasing the resources needed to implement the solution.

Benefits of using Shibboleth

- There has been international take-up, and SAML. is a de-facto standard
- “Out of the box” applications ready for Shibboleth (“shibd”) are available
- There is only one web login technology to support, meaning less administrative effort, less documentation and less user education is required
- There is less of a burden on web developers are they no longer have to understand how to set up secure logins and how or where to get user data

UK Access Management Federation
Shibboleth is a federated technology. As described above, the architecture separates authentication from authorisation. Authentication is controlled by the user’s home institution, while authorisation is based on the user’s role(s) within the institution and is controlled by the service provider (e.g. publisher).

A group of institutions and organisations that sign up to an agreed set of policies for exchanging information about users and resources to enable federated access and use of resources and services is called a federation. In November 2006, JISC and Becta launched the UK Access Management Federation 14. JISC has invited all UK HE/FE institutions to join the UK federation and adopt federated access management technologies.

Eduperson Standard
The eduPerson15 object class was developed by an EDUCAUSE/Internet2 working group. The aim of the group was to provide a set of directory attributes and definitions that can represent individuals in higher education. The group draws on the work of educational standards bodies in selecting definitions of these

---

15 [http://www.educause.edu/eduperson/949](http://www.educause.edu/eduperson/949)
directory attributes. These attribute definitions are added to the schema of an adopting institution's directory by importing an LDIF file containing the definitions of each eduPerson class and its attributes.

An initial scoping study[^16] for 'UKeduPerson' was commissioned by JISC in early 2004. Since the UK Access Management Federation was set up in 2006 it has taken its core attributes from the eduPerson schema.

**XACML**

OASIS eXtensible Access Control Mark-up Language[^17] (XACML) is a general-purpose access control policy language, providing a syntax (defined in XML) for managing access to resources.

The latest version 2.0 was ratified by OASIS standards organization on 1 February 2005. Version 3.0 is in preparation (since 2007) and will add generic attribute categories for the evaluation context and administrative policy profile. In education in the UK and elsewhere, experimentation with XACML has mostly been confined to the GRID community[^18].

XACML describes two parts of a particular application; the policy part used to describe general access control requirements; and the request/response part, which allows construction of a query to determine whether a particular action should be permitted.

XACML exploits long-established techniques, such as:

- Combining independent rules to form a single policy.
- Combining independent policies, from different policy-writers, to form a single policy set.
- Defining the set of decisions that the rule or policy is intended to render in terms of the name or attributes of the subject, resource and action identified in the decision request.
- Specifying in a policy statement a set of actions that must be performed in conjunction with the rendering of a decision.
- Providing an abstraction layer between the policy language and the environment to which it applies.
- The communication of policies, either attached to the resources they are intended to protect, or separately.

There is an increasing need for flexible management of protected content as part of institutional or other repositories, but most approaches currently rely on proprietary authorisation mechanisms that may be inefficient, costly, and inflexible, particularly in terms of sharing in a federation. As a practical example, the RAMP[^19] project has been addressing the need for open standards authorisation through the creation of a generalised XACML authorisation module that could potentially be adopted by any repository system.

**References and further reading**

- The Shibboleth project [http://shibboleth.internet2.edu](http://shibboleth.internet2.edu)

[^18]: e.g. in UK: Kent University, Chadwick, GRID use: [http://www.ogf.org/Public_Comment_Docs/Documents/2008-05/XACMLContextProfile0-5.pdf](http://www.ogf.org/Public_Comment_Docs/Documents/2008-05/XACMLContextProfile0-5.pdf)
- Secure Environment for Certificated Use of Resources (SECURe)
  http://www.angel.ac.uk/SECURe/

- UK Access Management Federation
  http://www.ukfederation.org.uk/

- JISC Access Management
  http://www.jisc.ac.uk/whatwedo/themes/access_management/federation.aspx

- JISC Access Management Glossary
  http://www.jisc.ac.uk/whatwedo/themes/accessmanagement/federation/glossary.aspx
Metadata

Metadata is a crucial part of any content digitisation or creation project and must not be neglected. Without metadata identifying its source, contents, and details of creation at an absolute minimum, a digital object is useless. Capturing additional information to facilitate rights management, administrative tracking, preservation, and distribution can enable more use from digital materials and add a high level of added value to the user experience.

An interesting approach to understanding metadata standards was drawn up by staff at the Université de Montréal in Canada, who created the MetaMap. Currently this requires SVG to run, and the Adobe website indicates that it will be removing support for this on 1 January 2009, so unfortunately this valuable resource may become an example of standards obsolescence. However, the project team hopes to migrate the resources.

© Université de Montréal

The acronyms on the map link to definitions and standard websites. The metro-map-inspired layout tries to provide a context for the standards, with coloured lines relating them to particular formats (e.g. still images, text), communities (e.g. archives, museums) or activities (e.g. creation, dissemination, preservation). It gives a visual overview of all the standards associated with metadata.

Who should provide metadata?
Traditionally libraries provided resources for cataloguing, and publishing companies added metadata to their publications. The huge growth in number of electronic resources, and the accompanying tendency for authors to publish their own material, means that this system has become unsustainable.

20 http://www.mapageweb.umontreal.ca/turner/meta/english/index.html (an SVG plug-in may be required)
Some metadata is provided by the content object itself: file type, file size, date created, and so on. Other metadata can be captured as part of the save/deposit/upload process, such as content title and author name. Further metadata that describes the content, its focus and contextual information, often needs to be input specifically. Such an approach is dependent upon the willingness and ability of staff to add this additional metadata. We are now more dependent on the willingness and ability of authors to add metadata according to accepted standards. Adding this extra burden to potential contributors may discourage them from contributing. It also runs the risk that, by removing professional input, metadata becomes less accurate and consistent and as a result less useful. We are then faced with the problem of how we check and validate author-generated metadata. Although there are technologies to enable semi-automated metadata generation, and thus lower the burden on the staff, the issues of checking and validating still apply.

The Metatools Project, being carried out at AHDS\(^\text{21}\), aims to develop a methodology for evaluating metadata generation tools; compare the quality of currently available metadata generation tools; and develop, test and disseminate prototype web services that integrate metadata generation tools. The project will base its investigation on three recent studies - the US AMeGA report\(^\text{22}\), which identifies the generic functionality required of metadata generation tools, JISC’s unpublished Metadata Generation for Resource Discovery study, which has identified key tools, gaps and areas for future research and development, and the JORUM Automated metadata\(^\text{23}\) report.

### Item-level and Collection-level Metadata

Item-level metadata applies to an individual item, and should be associated with and linked to the item it describes. Most metadata is applied to individual items.

There is no dominant metadata standard for describing collections, although in the last few years there has been substantial progress towards this goal\(^\text{24}\). NISO has also recently released a set of guidelines for building good digital collections\(^\text{25}\). UKOLN’s Collection Description Focus offer advice in this area, including a tutorial which points to the role of schemas\(^\text{26}\).

Some vital collection-level attributes resist conversion to item-level attributes, including metadata indicating that a collection is complete or incomplete, or was designed for some particular purpose\(^\text{27}\). These examples are tightly tied to the distinctive role a collection is intended to play. If this information is inaccessible, the collection cannot be useful, as a collection, in the way originally intended by its creators.

The Dublin Core Collections Applications profile\(^\text{28}\) (DCCAP) has been endorsed by DCMI so is effectively a standard. The NISO Metasearch Initiative’s Collection Description Specification (Z39.91) is based on DCCAP, which itself is based on an earlier de facto standard the RSLP Collection Description Schema\(^\text{29}\).

Various resources use DCCAP as the basis of their collection level metadata, including the IESR and The European Library (TEL).

Another collection level metadata standard is Encoded Archival Description\(^\text{30}\) (EAD), generally used in the archival community, and by the Archives Hub.

### Dublin Core

Dublin Core\(^\text{31}\) is one of the simplest and most widely used metadata schemas. If there is no other suitable schema for a collection, then Dublin Core is a good starting point.

---

\(^\text{21}\) [http://www.jisc.ac.uk/whatwedo/programmes/repres/tools/metatools](http://www.jisc.ac.uk/whatwedo/programmes/repres/tools/metatools)


\(^\text{24}\) [http://imlsdcc.grainger.uiuc.edu/resources.asp](http://imlsdcc.grainger.uiuc.edu/resources.asp)

\(^\text{25}\) [http://framework.niso.org/node/5](http://framework.niso.org/node/5)

\(^\text{26}\) [http://www.ukoln.ac.uk/cd-focus/cdfocus-tutorial/schemas/](http://www.ukoln.ac.uk/cd-focus/cdfocus-tutorial/schemas/)


\(^\text{28}\) [http://www.ukoln.ac.uk/metadata/rslp/schema/](http://www.ukoln.ac.uk/metadata/rslp/schema/)

\(^\text{29}\) [http://www.loc.gov/sard/](http://www.loc.gov/sard/)

\(^\text{30}\) [http://dublincore.org/](http://dublincore.org/)
Dublin Core has become a more generic standard and is now used to describe a wide range of different resources by many different communities, although it was originally intended as a way of describing Web sites. Dublin Core is one of the few metadata schemas that has achieved international standardisation (ISO15836). Because it is so simple and generic, the core metadata provided by Dublin Core is extended by the mechanism of Application Profiles to provide more specific and appropriate metadata for use in different communities and domains. It is also taking on an important role in interoperability since other more complex standards can usually be mapped to it.
Technical and Administrative Metadata

Technical and administrative metadata is used to facilitate management, tracking, migration and re-use of digital assets. It typically includes information on creation, quality control, rights and preservation. Some of the information may be harvested from the file itself while other information will need to be provided by the institution managing the image capture process. Administrative metadata can include critical information with respect to the conditions at the time of the digital capture; this may allow for some element of quality control.

Image metadata

DIG35 Metadata Standard for Digital Images

The overall goal of the DIG35 initiative was to define a standard set of metadata for digital images that will improve the semantic interoperability between devices, services and software. In August 2000 the DIG35 Metadata Specification\(^{32}\) was released providing a consistent set of metadata definitions to the imaging industry. The standard set of metadata can be widely implemented across multiple image file formats and provides a uniform underlying construct to support semantic interoperability of metadata between various digital imaging devices. The DIG35 Initiative also educates the industry regarding the importance of metadata usage, preservation and exchangeability.

VRA Core

Developed by the US Visual Resources Association, the VRA Core\(^{33}\) is a widely used metadata schema for describing images, particularly art or cultural images. VRA Core has been much influenced by the CDWA standard and by Dublin Core. The formal definition of VRA Core includes a mapping of its categories to Dublin Core.

The Images Application Profile project\(^{34}\) ran from September 2007 - April 2008. At this time (August 2008), the Application Profile is complete and the various project deliverables are detailed in the project wiki. VRA Core 4.0 was one of the models that the project looked at in detail. Work will be ongoing throughout 2008/9 to achieve a level of take up by the Repositories community. The project outlined a number of use cases outlining user needs, including the application of VRA Core 4.0.

MIX

NISO Metadata for Images in XML\(^{35}\) (MIX) is the generally accepted standard for still images. MIX is a version in XML of an extensive set of elements (the NISO Data Dictionary\(^{36}\) devised by the National Information Standards Organisation (NISO) for the detailed technical description of still images. The range of information that can be encoded in MIX is very large, from basic information on file types and sizes, to details of image capture (including capture hardware and image targets), to details of how an image has been processed after capture. Although a MIX file can be very lengthy and complex, almost all of its components (more than in its parent element set) are optional so that a basic record may be very simple. Although still in the process of drafting (version 0.2 currently), MIX has already established itself as the key standard for this type of metadata.

The NISO Data Dictionary has been designed to facilitate interoperability between systems, services, and software as well as to support the long-term management of and continuing access to digital image collections. The standard only applies to still raster (bitmap) images and does not address other image formats, such as vector or moving picture. The standard can be applied to digital images created through

\(^{32}\) http://xml.coverpages.org/FU-Berlin-DIG35-v10-Sept00.pdf
\(^{33}\) http://www.vraweb.org/vracore3.htm
\(^{34}\) http://www.ukoln.ac.uk/repositories/digirep/index/Images_Application_Profile
\(^{35}\) http://www.loc.gov/standards/mix/
\(^{36}\) http://www.niso.org/kst/reports/standards?step=2&gid=None&project_key=b897b0cf3e2ee526252d9f830207b3cc9f3b6c2c
digital photography or scanning, as well as those that have been altered through editing or migration (image transformation).

The Depot support service at EDINA will be applying MIX to provide technical and process metadata for still images. Information from TIFF file headers can be harvested and exported to relevant fields in XML output. Process metadata relating to the cropping & cleaning of images can also be recorded.

**Multimedia metadata**

For the technical and administrative metadata associated with audio and video files, there is no comparable standard to MIX. However, the following are in use:

**AUDIOMD**

The Audio Technical Metadata Extension Schema\(^{37}\) includes all key information necessary to make sense of an audio file (including, for instance, its format, bit rates, sampling frequencies, and any compression applied to it). The schema is of minimal size and complexity but functional.

**VIDEOMD**

Video Technical Metadata Extension Schema\(^{38}\) is a set of 36 elements designed for the Library of Congress's digital library projects. It concentrates solely on technical metadata and so avoids any potential problems of overlap with schemes for other types of metadata.

**PBCore**

A popular schema for video in the US is PBCore\(^{39}\), produced by public broadcast television services. This includes a set of elements for the technical description of the digital video file itself, including details of file formats, encoding, duration, aspect ratios and details of changes made as it is processed after creation. However, it also includes elements for descriptive metadata and intellectual property information, which may overlap with other schemas such as MODS or PREMIS.

**Structural metadata**

Structural metadata describes the internal structure of digital resources and the relationships between their parts. It is used to enable navigation and presentation.

**METS**

The METS schema is a de facto standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using an XML Schema. METS is an established standard for structuring complex digital resources (e.g. publications with multiple pages) and wrapping other sets of metadata. It is often used with MODS.

Use of METS has a number of benefits, which were outlined by the Paradigm\(^{40}\) project as follows:

- METS is maintained by the Library of Congress and is non-prorietary;
- Any system capable of handling XML documents can be used to create, store and deliver a METS file, thereby mitigating problems of software obsolescence;

\(^{37}\) [http://lcweb2.loc.gov/mets/Schemas/AMD.xsd](http://lcweb2.loc.gov/mets/Schemas/AMD.xsd)
\(^{38}\) [http://lcweb2.loc.gov/mets/Schemas/VMD.xsd](http://lcweb2.loc.gov/mets/Schemas/VMD.xsd)
\(^{39}\) [http://www.pbcore.org/](http://www.pbcore.org/)
\(^{40}\) [http://www.paradigm.ac.uk/index.html](http://www.paradigm.ac.uk/index.html)
- It is written in XML, which is robust as an archival medium; also it is an XML Schema extensible to future additions and supporting the use of multiple XML namespaces, which allow different kinds of metadata to be encoded in the same document;

- METS has the ability to deal with a wide variety of materials;
  - METS was designed to act as an OAIS (Open Archival Information System) Information Package, so can deal with all categories of OAIS metadata; it packages this metadata with the digital object it describes, ensuring that the object is self-documenting over time and it links together all the categories of metadata for an object, even if they are stored in separate locations;
  - METS effectively expresses the hierarchical structure of digital objects;
  - Whilst METS has the capability to deal with large and complex digital objects which might be comprised of many files, it is also useful for dealing with individual files (e.g. a single word-processed document). The latter is useful for digital archives, where long-term digital preservation requires that extensive metadata be recorded for each individual digital object;
  - The possibility of creating multiple structural maps in a METS document means that archivists can also take advantage of its capacity for sorting and reordering records in varied ways for researcher access.

However, METS also has a number of weaknesses, also outlined by the Paradigm project, and commented upon recently in the context of the British Library experience of using METS:

- The flexibility of METS can raise interoperability problems. It does not ensure standardisation because it does not operate as a metadata standard, rather as a framework within which metadata can be stored;

- Some metadata is difficult to use without bespoke development;

- Whilst using METS Profiles can mitigate these problems to an extent and facilitate manual cross-mapping, this still does not allow the automatic transfer of files between systems;

- At present METS documents largely have to be generated manually;

- METS relies on the effective use of unique identifiers. This can be difficult to administer.

**MPEG-DIDL**

Similar to METS, the de facto standard MPEG-DIDL (Digital Item Declaration Language) has also been proposed as a suitable vehicle to support transfer and dissemination of complex objects for preservation by an external service provider.

**Bibliographic, descriptive or delivery metadata**

Bibliographic, or descriptive, metadata describes and identifies information resources, to facilitate searching, retrieval, and management. It typically includes basic bibliographic information such as the creator, title, creation date; catalogue information such as accession or other identifying numbers; and topic information such as key words. Examples of descriptive metadata include Library of Congress Subject Headings, Categories for the Description of Works of Art (CDWA), the Art and Architecture Thesaurus (AAT), and the Dublin Core metadata set.

Access to digitised items via COPAC and the OPACs of holding libraries is enabled by adding a link to existing MODS, MARCXML or MARC21 records.

---

41 Reference Model for an OAIS [http://public.ccsds.org/publications/archive/650x0b1.pdf](http://public.ccsds.org/publications/archive/650x0b1.pdf)
44 [http://dublincore.org/](http://dublincore.org/)
The Categories for the Description of Works of Art (CDWA) is a standard for cataloging cultural objects, such as those found within museums and galleries. It has 512 categories or sub-categories, so it is more detailed and extensive than VRA Core. However, recognizing that such a lengthy standard is not usable for many institutions or collections, the CDWA identifies a set of 35 core categories which should be used as a minimum.

The MARC family of metadata is a standard format used by library systems to catalogue and exchange details of their holdings. It provides a standard machine-readable format in which to record metadata about bibliographic items. It is used extensively by libraries and bibliographic services such as COPAC. Variations of MARC that work with current technologies, including XML, are available; MARCXML is more extensive than MODS (see below). MARC is one of the formats available for results retrieved by searching technologies used within the JISC Information Environment.

The National Library of Medicine (NLM) standard is widely used for journals but has been adapted for monographs, such as pamphlets. JSTOR are implementing an adaptation of the NLM Journal Archiving and Interchange DTD. However, it has been found that the DTD is time consuming to implement, so it is unlikely to see wide take up.

Metadata Object Description Schema (MODS) is a schema for a bibliographic element set that can be used for a variety of purposes, and particularly for library applications. The standard is maintained by the Network Development and MARC Standards Office of the Library of Congress with input from users.

MODS was developed to hold a simplified set of MARC data for use within digital library collections. MODS uses a sub-set of MARC encoded as XML. In this XML form MODS is suitable for combining with other metadata. What is in several data elements in MARC may be brought together into one in MODS. For example, those digitising a book might choose to use MODS to describe the book as a whole, Dublin Core to describe the individual page image files, and METS to wrap the various records together. MODS is currently version 3.3 and has been adopted by COPAC.

The main advantage to using MODS is that it is intended to complement other metadata formats. For some applications, particularly those that have used MARC records, there will be advantages over other metadata schemes. Some other advantages are that the element set is richer than Dublin Core, simpler than the full MARC format, and compatible with library data; the schema is more end user oriented than the full MARCXML schema.

However, although as an element set for original resource description MODS allows for a simple record to be created, in some cases using more general tags than those available in the MARC record, there is no direct crosswalk: an original MARC 21 record converted to MODS may not convert back to MARC 21 in its entirety without some loss of specificity in tagging or loss of data. That said, the data itself will not be lost, only the detailed identification of the type of element it represents.

A further weakness is that MODS does not include rules for populating the elements. There are only comments to refer the user to where the semantics for the elements listed may be found.

---

45 see http://www.getty.edu/research/conducting_research/standards/cdwa/
46 MARC. http://www.ifla.org
47 COPAC. http://copac.ac.uk
48 http://dtd.nlm.nih.gov/archiving/
49 http://www.loc.gov/standards/mods/mods-overview.html
Spectrum
Standard ProcEdures for CollecTions Recording Used in Museums\textsuperscript{50} (SPECTRUM) is the international museum industry standard for collections management. SPECTRUM is now published by the Collections Trust as an open standard on behalf of the libraries, archives and museums sector, free of charge for non-commercial use. Spectrum is now in Edition 3. SPECTRUM provides definitions and guidance for a total of 21 procedures, or business processes involving collections, as well as identifying and describing the information which needs to be recorded to support the procedures.

SPECTRUM includes information on the minimum UK standard for museum documentation. A simplified version, called SPECTRUM Essentials, is available for smaller museums. SPECTRUM is a well-respected standard internationally, and is increasingly used as the basis for international interchange of museum data.

PRISM
The Publishing Requirements fo r Industry Standard Metadata\textsuperscript{51} (PRISM) specification defines an XML metadata vocabulary for managing, aggregating, post-processing, multi-purposing and aggregating magazine, news, catalog, book, and mainstream journal content. PRISM recommends the use of certain existing standards, such as XML, RDF, the Dublin Core, and various ISO specifications for locations, languages, and date/time formats. In addition PRISM provides a framework for the interchange and preservation of content and metadata, a collection of elements to describe that content, and a set of controlled vocabularies listing the values for those elements.

PRISM consists of two specifications. The PRISM Specification provides definition for the overall PRISM framework. A second specification, the PRISM Aggregator DTD is a new standard format for publishers to use for delivery of content to web sites and to aggregators and syndicators. It is an XML DTD that provides a simple, flexible model for transmitting content and PRISM metadata.

Application Profiles
The Scholarly Works Application Profile\textsuperscript{52}, originally called the 'Eprints Application Profile', is a Dublin Core Application Profile for describing scholarly works (eprints) held in institutional repositories. SWAP aims to help support richer searching and also to support full text indexing\textsuperscript{53}. SWAP is based on the Dublin Core Abstract Model allowing descriptions of different entities to be grouped together into a description set for sharing as a metadata record.

Further application profiles are being developed for Geo-spatial data/information, Images, and multi-media resources, and some further investigation into the development of application profiles for learning objects and scientific data has been carried out.

Preservation metadata

PREMIS
PREMIS is a de facto standard metadata schema for preservation metadata. PREMIS was developed by an OCLC/RLG working group and is being maintained via the Library of Congress\textsuperscript{54}. It was particularly influenced by the Open Archival Information System (OAIS), which provides a framework for the long-term preservation of digital (and non-digital) resources.

\textsuperscript{50}http://www.mda.org.uk/spectrum.htm
\textsuperscript{51}http://www.prismstandard.org/
\textsuperscript{52}http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_Profile
\textsuperscript{54}http://www.loc.gov/standards/premis
The PREMIS data dictionary specifies core metadata elements used to support the preservation of digital resources. It can be used for verifying and tracking the provenance, and checking the authenticity and integrity of preserved digital assets. It is being widely adopted as a means of recording information to support the preservation of digital resources. However, PREMIS is currently still undergoing a period of trial use.

The Paradigm project has undertaken a similar review of the strengths and weaknesses of PREMIS to that carried out for METS. Amongst others, they list the core strengths of PREMIS as follows:

- PREMIS allows references to external sources of information, e.g. format registries and policies.
- Has XML schemas available.
- Requires establishment of preservation policy to enable preservation level to be specified.
- Is based on the practical experience of many institutions
- Is flexible: there are no requirements as to how preservation information is stored.
- Is applicable to all types of digital material
- It can be incorporated into a METS file and therefore combined with other metadata to create an Information Package.

However, Paradigm also looked into the weaknesses of PREMIS, including difficulties in its implementation and requirements. They found that one of the main concepts on which PREMIS is based, representation, may be difficult for users to grasp. PREMIS requires local agreement on data content standards, which means work on selecting and, in many cases, developing controlled vocabularies based on repository policies which may not yet exist in many repositories. Further this may reduce its ease of interoperability, unless standards for these emerge.

Currently there are insufficient implementation examples, though more are emerging. Finally, PREMIS must be supplemented by metadata which can record detailed technical attributes of specific object-types, or media and hardware.

References and Further Reading (on Metadata)

- Understanding Metadata [http://www.w3.org/TR/UNDERSTANDING-WCAG20/appendixC.html](http://www.w3.org/TR/UNDERSTANDING-WCAG20/appendixC.html)
- Metadata and Digital Repositories [http://wiki.cetis.ac.uk/Metadata](http://wiki.cetis.ac.uk/Metadata)

TASI Metadata Standards and Interoperability [http://www.tasi.ac.uk/advice/delivering/metadata-standards.html](http://www.tasi.ac.uk/advice/delivering/metadata-standards.html)
Digitisation

What do we mean by digitisation?
Digitisation is the creation of a digital version of physical media. The physical objects will vary hugely. Digitisation includes the creation of images, rekeying, OCR, video and audio conversion, in theory directly from the original but this is not always possible or practical. Recommendations and standards for the creation of digital media together with the hardware and software used in the digitisation process are constantly changing. As with many continually developing processes cost also tends to fluctuate accordingly.

Why digitise?
The chief reasons for undertaking a digitisation project are usually given\(^{55}\) as:

- **Widening access:** The online publication of resources opens them up to a large potential audience, both in the UK and abroad; access to little used or unknown collections can be opened up;

- **Making new connections:** The juxtaposition of many different but related resources may reveal connections between resources that would be difficult or impossible to make; opening up a collection to a wider audience may enhance understanding of the collection and how it fits with other collections;

- **Preservation of resources:** Digitisation helps to preserve resources by creating a digital copy of the resource, and lessening the pressure of use on the physical object;

- **Providing a better understanding:** for original works this could be through improved indexing or some form of digital image enhancement.

The Technical Advisory Service for Images\(^{56}\) (TASI) provides further guidance on the digitisation process for images. A series of resources developed by the Arts and Humanities Data Service\(^{57}\) (AHDS) are still available and contain a further useful overview of many of the issues addressed in this document.

Preparation for digitisation
Previous guidelines written on the topic\(^{58}\) agree that project staff must have a good knowledge of the collections to be digitised and the uses that will be made of the digital resources created. Further, guidelines indicate that project staff should be aware of UK wide and international large-scale digitisation initiatives. Finally, project staff should also be aware of methods for cost reduction for the digitisation process such as outsourcing, automating digitisation and metadata creation, streamlining workflow, continuous improvement and quality assurance.

It is common for guidelines on digitisation to suggest that each object should be catalogued either before being digitised or during the project. The cataloguing should include essential metadata about the object, interpretation of the object, context of each object and what it links to, and how to find the original object. This information is essential for preservation and resource discovery purposes.

---

\(^{55}\) e.g. see: [http://www.tasi.ac.uk/advice/managing/decide.html](http://www.tasi.ac.uk/advice/managing/decide.html)
\(^{56}\) TASI [http://www.tasi.ac.uk/](http://www.tasi.ac.uk/)
\(^{57}\) [http://ahds.ac.uk/creating/guides/#visual-arts](http://ahds.ac.uk/creating/guides/#visual-arts)
\(^{58}\) Creating Digital Images, TASI, [http://www.tasi.ac.uk/advice/creating/creating.html](http://www.tasi.ac.uk/advice/creating/creating.html)

The Digitisation Process, UKOLN [http://www.ukoln.ac.uk/nof/support/help/papers/digitisation.htm](http://www.ukoln.ac.uk/nof/support/help/papers/digitisation.htm)

Capture and conversion

Hardware
The nature of the material to be digitised, particularly its condition and construction, determines the type of equipment used for capture. All collections should be surveyed prior to digitisation to decide on the best materials for digitisation. Equipment requiring minimal handling of objects during the digital capture process should be preferred.

When selecting digitisation hardware and software, project staff responsible must take into account the characteristics of the original objects such as format, size, condition and whether it is important to capture accurately attributes such as colour.

Software
The standards outlined in the following table are the image, moving picture and sound standards in general current use. As outlined in the UKOLN Standards Catalogue, there are certain other standards (e.g. BMP (Bitmap)) which are old, obsolete, or not generally used since they have been superseded by other standards.

<table>
<thead>
<tr>
<th>Standard and references</th>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>GIF (*.gif)</td>
<td>GIF is widely supported by a large number of viewing and editing programmes. GIF is an ideal graphics choice for drawings, cartoons, line art, icons and any image with large (horizontal) blocks of single colours, including animations.</td>
<td>GIF’s 8-bit (256) colour limitation produces a noticeable degradation in the quality of some images, especially photo-realistic images and images containing smooth colour gradients.</td>
</tr>
<tr>
<td>JPEG (*.jpg)</td>
<td>Widely supported format in terms of viewing and editing applications. JPEG compression gives very high compression rates with high perceived image quality and can handle high resolution images with colour depths of 24-bit and above. The newer progressive JPEG format allows for quicker initial display of images similar to GIF interlacing. JPEG is the dominant format for photo-realistic images on the Web.</td>
<td>JPEG has difficulty efficiently and effectively compressing images containing sharp edges. JPEG is a lossy compression method, i.e. it discards information each time it is used to compress an image, so images usually have to be stored in an intermediate format if they need further manipulation. There is no agreed standard to measure JPEG compression and each editing application has developed its own compression or quality scale.</td>
</tr>
<tr>
<td>JPEG2000 (*.jp2)</td>
<td>JPEG 2000 is a new image coding system that uses state-of-the-art compression techniques based on wavelet technology. Its architecture lends it to a wide range of uses from portable digital cameras through to advanced pre-press and medical imaging. JPEG2000 allows an image to be retained without any distortion or loss.</td>
<td>JPEG2000 is not widely supported in Web browsers and so is not generally used. JPEG2000 has higher computational and memory demands than JPEG; PNG is more space efficient and supports compression features that JPEG2000 does not. Wikipedia reports that there may be issues over the legal status of JPEG2000.</td>
</tr>
</tbody>
</table>

GIF (*gif)
Graphics Interchange Format is an established, proven colour graphics standard for generating highly compressed raster images
http://www.w3.org/Graphics/GIF/spec-gif89a.txt

JPEG (*.jpg)
Developed by the Joint Photographic Experts Group, this format is offered by most digital cameras
JPEG FAQ:
http://www.faqs.org/faqs/jpeg-faq/part1/preamble.html

JPEG2000 (*.jp2)
Developed by the Joint Photographic Experts Group committee in the year 2000 with the intention of superseding JPEG
http://www.jpeg.org/jpeg2000/
<table>
<thead>
<tr>
<th>Standard and references</th>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PNG (*.png)</strong>&lt;br&gt;Portable Network Graphics: Support for true colour images and variable transparency. <a href="http://www.libpng.org/pub/png/">http://www.libpng.org/pub/png/</a></td>
<td>PNG provides an improved means of display through 2D interlacing, offering up to 16-bits, and lossless compression. It is a good format choice for storing intermediate-stage images; it could replace GIF. The source code required to read and write PNG files is free. PNG is a W3C recommendation for bitmap graphics.</td>
<td>Although PNG is a very powerful format, it does not support animation. PNG is not as widely supported by Web browsers as GIF. PNG does not provide a standard means of embedding Exif image data from sources such as digital cameras.</td>
</tr>
<tr>
<td><strong>SVG (*.svg)</strong>&lt;br&gt;Scalable Vector Graphics <a href="http://www.w3.org/Graphics/SVG/">http://www.w3.org/Graphics/SVG/</a>&lt;br&gt;Case Study: Sustainability of Digital Formats, Planning for Library of Congress Collections: <a href="http://www.digitalpreservation.gov/formats/fdd/fdd000020.shtml">http://www.digitalpreservation.gov/formats/fdd/fdd000020.shtml</a></td>
<td>SVG is an open format with significant support from industry (including Apple, Microsoft, Sun Microsystems, Adobe). SVG is a W3C recommendation for vector graphics.</td>
<td>Support for SVG in older web browsers is poor; support in current browsers is incomplete. Take-up has not been widespread. Microsoft has no SVG support except via an unsupported plug-in.</td>
</tr>
<tr>
<td><strong>TIFF (*.tif)</strong>&lt;br&gt;Aldus Corporation designed and made public the Tagged Image File Format (TIFF) in 1986. Although targeted at desktop publishing, it has been widely implemented on many computing platforms and has become a de-facto industry standard format. <a href="http://partners.adobe.com/public/developer/tiff">http://partners.adobe.com/public/developer/tiff</a></td>
<td>TIFF is a tag-based format for storing and interchanging raster images. Typically TIFF describes data that comes from scanners, frame grabbers and paint programs. It is capable of describing bi-level, grayscale, palette-colour and full-colour image data in several colour spaces. There are four basic TIFF internal formats: bitmap, grey-scale, palette-colour and RGB-colour.</td>
<td>Even though TIFF is widely used there are problems with different versions since people can add their own tags which other users do not necessarily understand. It is not appropriate to use TIF format on websites.</td>
</tr>
<tr>
<td><strong>SWF (Flash) (*.swf)</strong>&lt;br&gt;A proprietary format, owned by Adobe. A display format as opposed to a capture format. <a href="http://www.techdis.ac.uk/index.php?p=3_7_26_2">http://www.techdis.ac.uk/index.php?p=3_7_26_2</a>&lt;br&gt;case study: <a href="http://www.ukoln.ac.uk/qa-focus/documents/case-studies/case-study-07/">http://www.ukoln.ac.uk/qa-focus/documents/case-studies/case-study-07/</a></td>
<td>Flash is the industry standard for displaying vector graphics and animation on the Web, and is increasingly used for Web video and rich interactive applications (RIAs). Since May 2008 SWF is no longer licenced, it has been released as part of Adobe’s Open Screen Project.</td>
<td>The newest version of the Flash player will normally need to be downloaded and installed to ensure that graphics can be viewed. Questions surround the usability of Flash, accessibility problems for those with disabilities, compatibility with certain platforms, and privacy concerns due to the use of cookies.</td>
</tr>
<tr>
<td><strong>MPEG (*.mpg)</strong>&lt;br&gt;MPEG provides an open, standard format for video (and audio) data compression <a href="http://www.chiariglione.org/mpeg/index.htm">http://www.chiariglione.org/mpeg/index.htm</a></td>
<td>MPEG is capable of producing high-quality video at smaller file sizes than comparable video formats. Good for distribution.</td>
<td>Lossy intra-frame compression in MPEG means that the data removed cannot be recovered. Native MPEG support is not present in older versions of Web browsers. For non-supported browsers, MPEGs have to be played through a plug-in. Unsuitable for preservation.</td>
</tr>
<tr>
<td>Standard and references</td>
<td>Strengths</td>
<td>Weaknesses</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-----------</td>
<td>------------</td>
</tr>
<tr>
<td>AVI (*.avi)</td>
<td>AVI is the most popular format for audio/video data and is widely supported on all operating systems.</td>
<td>Although AVI can be played on other platforms, it is a Microsoft proprietary format. AVI is of relatively low quality (e.g. when compared to MPEG-1) with a limiting frame rate and mono sound. Playback is dependent on the availability of 3rd party codecs.</td>
</tr>
<tr>
<td>Audio Video Interleave</td>
<td>QuickTime is a truly multimedia format which allows the integration of text, graphics, video, animation, 3D, VR and sound into one file. It is provided free of charge. ISO’s decision to base the MPEG-4 industry standard on QuickTime added support for an already popular format</td>
<td>An external (free) plug-in is required to view QuickTime files on Macintosh and Windows platforms. There is no plug-in support for Unix. Not suitable for preservation</td>
</tr>
<tr>
<td>QuickTime (*.mov)</td>
<td>RealVideo is a well established format, with a large user base. The latest playing software allows the user to fine tune the video in real time and has support for third-party and standard data types such as AVI, WAV, MIDI, MPEG and JPEG.</td>
<td>RealVideo files require RealServer server software (a basic version is free) to handle the outgoing streams of packets and RealPlayer software to receive and play the data on the client machine. The RealPlayer tends to be memory-intensive. Not suitable for preservation</td>
</tr>
<tr>
<td>A proprietary video format developed by RealNetworks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RealVideo (*.ram)</td>
<td>AU (*.au)</td>
<td>AU files can be relatively low quality and are not highly compressed. Not widely supported, so not suitable for preservation.</td>
</tr>
<tr>
<td>A well established format for the Java programming language. A simple, well-established sound format which is the most commonly supported browser sound file format. Mostly used on Unix machines.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AU is short for AUdio</td>
<td>WAV (*.wav)</td>
<td>Uncompressed WAV files are large in size, so, as file sharing over the Internet has become popular, use of the WAV format has declined.</td>
</tr>
<tr>
<td>The WAV, short for Waveform audio, file format was developed by IBM and Microsoft as the RIFF (Resource Interchange File Format) WAVE to store sound data in structured chunks.</td>
<td>WAV files are simple and widely used. Many applications have been developed to play WAV files and it is the native sound format for Windows. It is a commonly used, lossless file type, suitable for retaining “first generation” archived files of high quality, or use on a system where high fidelity sound is required and disk space is not restricted.</td>
<td></td>
</tr>
<tr>
<td>AIFF (*.aiff)</td>
<td>A well-used lossless format on Apple computers.</td>
<td>AIFF creates large files and is not used within some specialist communities</td>
</tr>
<tr>
<td>Audio Interchange File Format</td>
<td>OGG is a free, open standard format designed to provide for efficient streaming and manipulation of high quality digital multimedia. It has been adopted by some portable and proprietary media players.</td>
<td>Open standards might not be the most useful. It is not feasible for some kinds of format e.g. OGG for music rather than WAV or MP3 as the latter are more deeply embedded into current practice.</td>
</tr>
<tr>
<td>OGG (*.ogg)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><a href="http://www.vorbis.com/">http://www.vorbis.com/</a></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Standard and references

<table>
<thead>
<tr>
<th>Standard and references</th>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>**MIDI (<em>.midi, <em>.mid)</em></em></td>
<td>The MIDI standard allows for musical pieces to exchanged and edited by different computers on different platforms in a way that conventional digitised sound (actual waveforms) cannot. MIDI produces much smaller file sizes for music, compared to other sound formats. It has been widely adopted.</td>
<td>Browsers require separate plug-ins or helper applications to play MIDI files. The MIDI information is not able to guarantee an accurate waveform will be heard by the intended listener, because each MIDI synthesiser will have its own methods for producing the sound from the MIDI instructions provided. Unsuitable for preservation.</td>
</tr>
</tbody>
</table>

### Geographic Information Standards (GIS)

There are over 40 separate standards in the ISO geographic information series[^59], all aimed at standardisation in the field of digital geographic information. The standards support the understanding and usage of geographic information. GIS should also increase the availability, access, integration, and sharing of geographic information, i.e. to enable interoperability of geospatially enabled computer systems and data. Some of the standards are relatively mature (i.e. longitude and latitude measures), others are still in development (i.e. metadata).

Note that GIS can also stand for geographic information system, which integrates hardware, software, and data for capturing, managing, analysing, and displaying all forms of geographically referenced information. Typically, a GIS is used for handling maps of one kind or another. These might be represented as several different layers where each layer holds data about a particular kind of feature (e.g. rivers, roads, geology, archaeological sites). Each feature is linked to a position on the graphical image of a map. As well as being spatially referenced, layers can also hold any number of other attributes, for example a layer relating to archaeological sites could include a name plus site type and period attributes for each individual site.

### Mark-up

Almost all documents presented on the Internet have been tagged for presentation in either Standard Generalised Mark-up Language (SGML) or Hypertext Mark-up Language (HTML). These were standardised in the late 1980’s and early 1990’s. Extensible Mark-up Language (XML), a more flexible form of mark up language, was developed from SGML in the late 1990’s as it became clear that a standard for general Internet usability and ease of authoring was required following the rapid expansion of the Internet.

#### HTML, SHTML

One of the main strengths of HTML is the wide availability of applications, often for no charge, which can create and render HTML documents. The ease of transfer of HTML documents over the Web using the HyperText Transfer Protocol (HTTP), combined with a standard means of locating files by the Universal Resource Locator (URL), has given users access to a powerful, interactive global information resource.

However, the continuing development of vendor-specific tags and the proliferation of the use of HTML as presentational, rather than semantic, mark-up are factors which potentially undermine the original concept of HTML. Other weaknesses include a lack of internal structure in HTML and the inability to define data input fields as required. There is also a need to ensure that documents using HTML comply with relevant accessibility standards.

[^59]: For up-to-date status: [http://www.isotc211.org/](http://www.isotc211.org/)
SGML
SGML is a rich and flexible language of great value in managing large documents which need to be frequently updated and capable of being printed in different formats. SGML provides a very stable means of information exchange and is a reliable, industry-approved format for archiving and preservation.

In spite of these strengths, the fact that SGML is a very comprehensive language means that it is relatively complex and this has curtailed its adoption in many environments. Browser support is a potentially limiting factor, especially when compared to the availability of HTML browsers.

XML
XML began life in the 1960s as SGML (Standard Generalised Mark-up Language), a system for tagging up electronic texts using semantically meaningful tags. XML is far more extensible than SGML or HTML. Rather than providing a set of pre-defined tags, as in the case of HTML, XML specifies the standards with which you can define your own mark-up languages with their own sets of tags. XML is therefore a meta-mark-up language, allowing you to define an infinite number of mark-up languages based upon the standards defined by XML.

XML was created so that richly structured documents could be used over the web. The only viable alternatives, HTML and SGML, are not practical for this purpose. XML allows you to define all sorts of tags with all sorts of rules, such as tags representing business rules or tags representing data description or data relationships.

All of the standards discussed here use XML as their semantic and structural underpinning. In addition to marking up texts themselves, it has also come to be used increasingly as a standalone mechanism for encoding metadata for all types of objects in traditional or electronic libraries. XML incorporates a greater degree of functionality within documents than HTML, but retains the ease with which documents can be served and processed. The strengths of XML as the basis of a metadata scheme have often been acknowledged. It is a fully open standard registered with the ISO (International Standards Organisation), and so is independent of any given software application.

Although use of XML is now ubiquitous, as with all standards, developers have noted some drawbacks and weaknesses of XML, including:

- The verbose nature of some XML mark-up, depending on the vocabulary in question; which may also lead to high storage costs;
- XML has difficulties dealing with multiple hierarchies;
- XSL/XSLT differences occur across browsers, so transformations cannot be relied upon;
- There is limited native browser support for XML, although plug-ins can be used to view XML documents;
- XML namespaces are problematic to use and namespace support can be difficult to correctly implement in an XML parser, parsers rarely combine completeness, correctness and efficiency;
- XML Hypertext Transfer Protocol (XML-HTTP) problems still exist.

DTD
An XML Document Type Definition (DTD) provides a description (actually expressed in the SGML DTD Specification language) of the building blocks of any type of XML document, whether that document is a list,
a metadata record, a journal article, or a whole book. It includes what to call different types of elements, how they should be ordered, and how they interrelate. Some DTDs are proprietary, for example, written by publishers for their own purposes\textsuperscript{68}, while others are standardised and freely available.

**XML schema**

An XML schema (also called an XSD file) is itself an XML document and is an alternative to the DTD that provides developers with enhanced validation capabilities and more refined tools for structuring their own XML-based formats. Whereas DTDs only allow for relatively simple data types, a schema has a set of powerful, flexible semantics for defining what an XML file can contain.

**XHTML**

Extensible HTML\textsuperscript{69} (XHTML) was developed to reproduce, subset and extend HTML4 documents. Strict XHTML mark-up is used to make web documents accessible, usable and consistent across web browsers. XHTML documents can utilise applications (e.g. scripts and applets) that rely on either the HTML Document Object Model or the XML Document Object Model, giving content forward and backward compatibility. XHTML 1.0 became a W3C Recommendation on 26 January 2000. W3C defines XHTML as the latest version of HTML. It is predicted that XHTML will gradually replace HTML.

**CSS**

CSS is used for rendering a document’s display. This is both to separate content from display information and to make a document usable and accessible to all users.

**Page presentation mark-up**

**DDI and TEI**

The Data Documentation Initiative\textsuperscript{70} (DDI) and the Text Encoding Initiative Guidelines (TEI) both map onto Dublin Core but provide more relevant detailed description.

The Data Documentation Initiative is an international effort to establish a standard for technical documentation describing social science data. DDI is in version 3.0\textsuperscript{71}, which represents an expansion of scope from previous versions, as it now includes not only data archiving but all aspects of the data lifecycle, from development to reuse. In constructing DDI 3.0 care was taken to review related standards to provide clear mapping to the contents of outside standards or to incorporate content where appropriate. DDI 3.0 currently has mapped relationships to the following standards: Dublin Core (Basic Bibliographic Information), MARC (Bibliographic Information), ISO/IES 11179 Data Registry, ISO 19118 (Geography), SDMX (Aggregate data), METS (Content Wrapper), PREMIS (Preservation).

The Text Encoding Initiative is an international consortium. TEI have developed an interdisciplinary standard which has established a set of models and guidelines for encoding texts in the humanities, including the representation of literary and linguistic texts online. TEI is the most widely used and longest established standard in this area. TEI enables the full electronic text of a book to be encoded along with supplementary metadata in a non-proprietary and open standard form. TEI also provides a rich and extensible set of mark-up tags, enabling the mark up of parts of the books for intelligent searching for textual analysis and retrieval of content. Where this mark-up is used, users can search for lines of verse, names, geographical locations, chapter headings and so on. TEI also offers the flexibility to present texts with minimal mark-up and then add more mark-up at a later date to add value to the text.

---

\textsuperscript{67} [http://www.w3.org/TR/REC-xml/](http://www.w3.org/TR/REC-xml/)
\textsuperscript{68} [http://www.diglib.org/preserve/hadtdfs.pdf](http://www.diglib.org/preserve/hadtdfs.pdf)
\textsuperscript{69} [http://www.w3.org/TR/xhtml1/](http://www.w3.org/TR/xhtml1/)
\textsuperscript{70} [http://www.ddialliance.org/](http://www.ddialliance.org/)
However, although TEI is good for scholarly works, it may not be appropriate for other types of user who want more simple types of mark-up. TEI Lite has been widely adopted, particularly by beginners and by big institutional projects that rely on large teams of encoders to markup their documents. A second less comprehensive, but simpler, option that is adequate for textual objects is a Schema for Technical Metadata for Text created by New York University; this is a set of 16 elements that provides all the information necessary for the rendering and display of texts in most contexts.

**PDF**

PDF is a device and resolution-independent means of efficiently transferring richly formatted documents between computers and across platforms. It is a well-established standard which has been widely adopted by publishers for on-line distribution of journals, manuals and books. Creators of PDF documents can incorporate various security features into the document to restrict access.

However, viewing PDF files with Web browsers requires an external helper application or plug-in. PDF files are not easily reformatted or edited without the appropriate Adobe-produced tool. As a proprietary format PDF is vulnerable to the whims of Adobe. Perhaps more importantly, PDF files have been noted to reduce website usability substantially.

**PDF/A**

PDF/A defines a format for the long-term archiving of electronic documents. It is based on the PDF Reference Version 1.4 from Adobe. PDF/A is a subset of PDF, which leaves out PDF features not suited to long-term archiving. See Digital Preservation section of the JISC Standards Document for more discussion.

**References and Further Reading**

- Creating Digital Images, TASI, [http://www.tasi.ac.uk/advice/creating/creating.html](http://www.tasi.ac.uk/advice/creating/creating.html)
- The Digitisation Process, UKOLN, [http://www.ukoln.ac.uk/nof/support/help/papers/digitisation.htm](http://www.ukoln.ac.uk/nof/support/help/papers/digitisation.htm)

---

73 [http://dlib.nyu.edu/METS/textmd.htm](http://dlib.nyu.edu/METS/textmd.htm)
74 Jacob Nielesen’s Alertbox column [http://www.useit.com/alertbox/20030714.html](http://www.useit.com/alertbox/20030714.html)
Digital Curation and Preservation

The Digital Curation Centre’s current definition is: *Digital curation, broadly interpreted, is about maintaining and adding value to, a trusted body of digital information for current and future use.*

The e-Science Curation Report\(^{75}\) proposed the following distinctions:

- **Curation**: The activity of, managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Higher levels of curation will also involve maintaining links with annotation and with other published materials.

- **Archiving**: A curation activity which ensures that data is properly selected, stored, can be accessed and that its logical and physical integrity is maintained over time, including security and authenticity.

- **Preservation**: An activity within archiving in which specific items of data are maintained over time so that they can still be accessed and understood through changes in technology.

Digital preservation is the set of processes and activities that ensure continued access to information and all kinds of resources existing in digital formats. This includes the preservation of materials resulting from digitisation, as well as information that is born-digital and has no physical counterpart. Preservation needs to be an ongoing process. The long-term storage of digital information is assisted by the inclusion of preservation metadata.

The Digital Curation Centre notes\(^{76}\) that there has been a lack of clarity surrounding the concepts of digital curation and preservation, and that this is partly due to a number of terms which have more or less clear definition being used too casually, either alone or in combination. The danger is that the reader (or listener) may believe that something extra is implied by the use of particular terms. DCC distinguishes the following terms:

- **data preservation**: a general term equivalent to digital preservation in this context

- **digital preservation**: sometimes interpreted as simply ensuring the original bits and bytes are accessible.

- **digital information preservation**: this is what is referred to in the OAIS standard - what is important is not the original "bits and bytes" but the content. An OAIS ensures that the content is accessible, understandable and usable.

- **curation**: this is a general term meaning to take care of things

- **data curation**: looking after and adding value to data

- **digital curation**: looking after and somehow "adding value" to digital data, ensuring its current and future usefulness. This probably implies creating some new data from the existing, in order to make the latter more useful and "fit for purpose".

Overall, the goal of digital preservation is to maintain the ability to display, retrieve, and use digital resources in the face of changing technological and organisational infrastructure. Issues to be addressed in digital preservation include:

- Retaining the physical reliability of the files, accompanying metadata, scripts, and programs (e.g., make sure that the storage medium is reliable with back-ups, maintain the necessary hardware and software infrastructure to store and provide access to the collection);

- Maintaining collection security (e.g., implement strategies to control unauthorised alteration to the collection, develop and maintain a rights management programme);

---

\(^{75}\) [http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf](http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf)

\(^{76}\) [http://twiki.dcc.rl.ac.uk/bin/view/Main/DCCApproachToCuration](http://twiki.dcc.rl.ac.uk/bin/view/Main/DCCApproachToCuration)
Ensuring continued usability of the digital collection (e.g. maintain an up-to-date user interface, enable users to retrieve and manipulate information to meet their needs).

For all digital objects the ability to persistently and uniquely identify each component is essential to ensure that objects can be correctly accessed and made available for end-user services. Ideally, identifiers should be location independent; that is they should not have links with the domain the content currently sits in. This offers a degree of future-proofing in case content and/or repositories of content move location, providing confidence that content referenced in metadata records will be available through resolution of the identifier in the long-term.

**OAIS**

OAIS\(^77\), the Reference Model for an Open Archival Information System, was developed by the Consultative Committee for Space Data Systems (CCSDS) to provide a framework for the standardisation of long-term preservation within the space science community. OAIS was created with a view to its being widely applicable to long-term preservation in any context, primarily, but not exclusively, digital. The model exists at an abstract level, providing a conceptual framework for raising “fundamental questions regarding the long-term preservation of digital materials – questions that cut across domain-specific implementations”\(^78\).

An evaluation of OAIS\(^79\), carried out by UKOLN in 2006, concluded that its applicability goes beyond the community it was developed for. The report went on to state that the conceptual and flexible nature of OAIS allows repositories to adapt and extend their own functional and informational models to take local practices into account whilst staying true to the OAIS model. In its current form, the length and apparent complexity of the OAIS document might prove a barrier to smaller repositories but a lightweight OAIS could be developed, offering a flexible, contextualised approach to applying the model across different types of repositories.

The OAIS Information Model has been used to inform the development of preservation metadata schemas, such as PREMIS, and draft schemas developed by the National Library of Australia and the Cedars project. METS objects are an implementation of OAIS information packages. The OCLC Digital Archive Service\(^80\) is based on OAIS.

**Preservation Related Standards**

Many digital preservation related standards are mentioned elsewhere in this document, as they fulfil another purpose as well as being useful for digital preservation. For example, the TIFF image format is very widely implemented in image capture, image processing and document management systems. It is not only used by scanners and digital cameras as a digital image interchange format but has also been adopted by some application developers as a memory format\(^81\).

The Digital Curation Centre’s DIFFUSE\(^82\) project aims to offer subject specific advice on standards relevant to digital preservation and curation, to help curators identify which standards they should be using and where they can be appropriately implemented. The forthcoming DCC Digital Curation Lifecycle Model will help to contextualise standards frameworks and relevant standards for a number of disciplines by presenting this information in a simple, visually descriptive format that can be utilised by a wide range of stakeholders. DIFFUSE Standards Frameworks are sets of standards which can be used together to achieve effective implementation of all stages of the Digital Curation Lifecycle. A standards framework will ideally support interoperability, maximise accessibility, avoid vendor lock-in, provide architectural integrity, and help to ensure long-term preservation.

\(^77\) OAIS Reference Model: http://www.ccsds.org/documents/650x0b1.pdf
\(^80\) http://www.oclc.org/digitalarchive/
\(^81\) http://www.dcc.ac.uk/diffuse/?l=1#52
\(^82\) http://www.dcc.ac.uk/diffuse/
Persistent Identifiers

Immense amounts of digital objects are now being produced and stored by research and education institutions. There is an obvious need to manage, discover and access these resources over time. Identifiers help manage resource repositories by separating the identification of resources (their names) from the location of resources (their addresses). Persistence of access to online digital materials has two aspects:

- Making sure that valuable online digital materials continue to exist;
- Making sure that they can be identified and accessed.

Persistent identifiers are a tool that can play an important role in identification and access. A persistent identifier is a name for a resource which will remain the same regardless of where the resource is located or whatever protocol is used to access it. It will never be reassigned to any other resource. Therefore links to the resource will continue to work even if it is moved. Persistent identifiers can relate to web resources (e.g. content in an institutional repository), e-research and e-science resources (e.g. instruments, computational services, data sets), or semantic web abstractions (e.g. concepts in an ontology).

The JISC IE technical standards document states: “Every significant item that is made available through a JISC IE network service should be assigned a URI that is reasonably persistent. This means that item URIs should not be expected to break for a period of 10-15 years after they have first been used. For this reason, JISC IE service components should not hardcode file format, server technology, service organisational structure or other information that is likely to change over a 10-15 year period into item URIs.”

The Digital Object Identifiers for Publishing and the e-Learning Community report83, from a study undertaken by the Stationary Office, concluded that persistent digital identifiers adopted by publishers in the education sectors in the UK should be able to:

- Reference multiple object types
- Integrate and interoperate with existing standards
- Offer scope for future extension and migration
- Satisfy the needs of the broad JISC community and other sectors
- Adopt declared IETF specifications or those under known IETF approval process
- Avoid semantic or location information in the identifier

Dack (200184) outlined three important practical aspects to persistent identification of online resources:

- The choice of an identifier system (What do we call the resources?);
- A system of resolution to map the identifier to the resource identified (How will users get the resources when the identifiers are used as links?);
- Maintenance of access to the resource through continued association of the current location of the resource with the identifier (How do we make sure that the links continue to work over time?).

Tonkin (200885) has recently published an excellent overview of the options surrounding the use of persistent identifiers. She outlines the different standards, discusses some of the differences between them, and gives some research issues and open questions to be answered by the community. This excellent paper will make a good resource for projects wishing to adopt a persistent identifier system. A similar short summary is given of the most commonly used persistent identifiers below.

Persistent identifiers are not unique to the web. Many formal identifier or naming schemes are used every day by various systems, either as internal system numbers or as public identifiers (e.g. ISBN). Similarly, where an ISBN is lacking, National Bibliography Numbers (NBNs) are a URN namespace used solely by

---

83 http://www.jisc.ac.uk/uploaded_documents/DOI%20JISC%20report.pdf
national libraries, in order to identify deposited publications which lack an identifier, or to reference
descriptive metadata (cataloguing) that describe the resources. There have been a number of efforts to
address the need for reliable persistent identifiers. A number of identifier schemes that have been developed
specifically for use on the Internet are at various stages of implementation and use. These are:

**URI/URL/URN**

Uniform Resource Locators (URLs) and Uniform Resource Names (URNs) are both types of Uniform
Resource Identifiers (URIs). A URI may be a name and a locator at the same time. The predominantly
location based use of the "http" URI that is its major weakness as a persistent identifier. It is also impossible
to tell whether it is being used as a name or an address. Uniform Resource Names (URNs) are intended to
serve as persistent, location-independent resource identifiers and are designed to make it easy to map other
namespaces (that share the properties of URNs) into URN-space.

**PURL**

Persistent URLs were developed by OCLC in the mid-1990s, primarily to reduce the maintenance burden
of the URLs contained in catalogue records created for Internet resources. Functionally, a PURL is a URL.
However, instead of addressing directly the location of an Internet resource, a PURL addresses an
intermediate resolution service or PURL server. This PURL Resolution Service associates the PURL with the
actual URL and returns that URL to the client software. The client then completes the URL transaction to
obtain the resource. The transaction is a standard Hypertext Transfer Protocol (HTTP) redirect, thus no
acceptance of new protocols or modifications to client software are required. The PURL software also
provides a convenient administrative interface for centrally updating indirect URLs. PURL is considered an
interim measure until URNs are more widely accepted.

**DOI**

A digital object identifier (DOI) is a permanent identifier given to an electronic resource. Similar to a URN,
but, in contrast to a URL, it is not dependent upon the electronic document's location.

The Digital Object Identifier System is a managed system for persistent identification of content on digital
networks. It can be used to identify physical, digital, or abstract entities; these names resolve to data
specified by the registrant, and use an extensible metadata model to associate descriptive and other
elements of data with the DOI Name. The DOI System is an implementation of URI using the Handle
System, which provides identifiers which are persistent, unique, resolvable and interoperable. It is useful for
automated management of content on digital networks. Web browsers can be adapted to support this
scheme with a plug-in. However, since the DOI system uses its own scheme, and supports an http alias for
each resource also, a significant disadvantage could arise in having to manage the two different names (for

The DOI System is being standardised through NISO; it is expected that the final standard will be published
in late 2008 or 2009. NISO have agreed the Digital Object Identifier (DOI) syntax as an open standard

**Handle**

The Handle System is a general-purpose global name service that allows secured name resolution and
administration over networks such as the Internet. It is not a standard as such, but an open set of protocols.
The Handle System manages handles, which are unique names for digital objects and other Internet

---

88 http://www.purl.org/
89 http://www.handle.net/
resources. The Handle System is infrastructure, not end user software. The current software release, HANDLE.NET 6.2, is written in Java™.

The Handle System provides services for allocating, managing and resolving identifiers. The resolution service resolves an object’s identifier with the actual location (web address) of that object. The PILIN92 project recommended use of the Handle System due to its free availability, its use in other systems such as DOI and DSpace93 (an open source repository), and that it can be used with legacy systems.

ARK
An Archival Resource Key94 is a URL created to allow persistent, long-term access to information objects. ARKs can identify objects of any type: digital documents, databases, images, software, and websites, as well as physical objects (books, sculpture, etc.) and even intangible objects (chemicals, diseases, performances). An ARK provides extra services beyond that of an ordinary URL. Instead of connecting to one thing, an ARK should connect to the object, a brief metadata, and a maintenance commitment from the current server.

The California Digital Library is currently adopting ARK for the objects that it owns or controls. Currently about 80,000 ARKs have been assigned at the CDL. CDL has found that because of the tight links between an ARK and its metadata, it is always easy to create bibliographies and object-level records that place semantically opaque ARKs right next to metadata that is not only fully expressive and authoritative, but that can also be refreshed automatically at any time. This means that ARK is not subject to the pressures forcing other identifier schemes (e.g. URI95 and Handle96) to deal with the serious complications surrounding inclusion of language and character set indicators inside identifiers.

Multiple systems of identifiers
The Digital Object Identifiers for Publishing and the e-Learning Community report97 recommended that multiple systems may be required. Therefore, the JISC would need to generate policy and interoperation specifications to ensure that it can interact with the range of existing identifiers. This would enable interoperable resource discovery, the exchange of metadata descriptions and digital objects, and aggregation relationships.

The whole JISC community would be unlikely to be able to use a single persistent identifier convention for the following reasons:

- A single persistent identifier implementation may restrict interoperability with other sectors.
- Many information objects or metadata records, in particular those that are provided by third party publishers, may already have persistent identifiers assigned to them;
- Informal sharing of information resources is likely to have different digital identifier requirements from that of the more formal traditional publishing.

The report concluded that there is a need to provide, and assure the continued availability of, more informal methods of creating persistent digital identifiers, which have low cost and minimal barriers for information providers.

92 Persistent Identifier and Linking Infrastructure (PILIN) Project: https://www.pilin.net.au; The Resourcing Identifier Interoperability for Repositories (RIDIR) project at Hull is a parallel project in the UK: http://www.hull.ac.uk/ridir/
93 http://www.dspace.org/
94 http://www.cdlib.org/inside/diglib/ark/arkspec.html
97 http://www.jisc.ac.uk/uploaded_documents/DOI%20JISC%20report.pdf
Questions and issues
From the perspective of a project manager looking at persistent identifiers for the first time, this collection of standards could be overwhelming. Each was developed to respond to different needs, although there are a number of common threads that link these initiatives together.

Tonkin (2008) suggests that choosing from these options on the basis of a brief written description is difficult if not impossible. Tonkin suggests that the choice of a persistent identifier system should start with requirements gathering and analysis. In common with other standards, the context in which the system will be used is an important factor, particularly since adoption of the persistent identifier standards is quite uneven and very dependent on the intended usage and topic area.

Tonkin gives a practical example, as particular attention has been put into a number of user scenario-related issues for the ARK; for example, practical issues such as the difficulties encountered copying a URL from paper into a browser location bar have been incorporated into the design. For this reason, Tonkin recommends the ARK specification and documentation for those looking into the various options for a persistent identifier standard. However, it should be noted that the standard that has achieved widest market penetration is likely to be the URN/DOI concept and its associated resolver.

The RIDIR (Resourcing IDentifier Interoperability for Repositories) project is investigating the requirements for, and benefits of, the clear use of persistent identifiers in order to facilitate interoperability between digital repositories of different types. As the primary mechanism for the management and dissemination of content resources migrates from the physical to the digital environment, the challenges for coherent models of resource identity become more pressing. There is an extensive discussion of many of the issues involved in the introduction to the RIVER project report. In summary, when it is easy to create and disseminate copies of resources, the exact identity of those resources becomes critical for users in many different contexts – although the ones with which the RIDIR project was particularly concerned were those within the discovery to delivery chain. RIDIR also commented that the importance of interoperability between identifiers in supporting preservation is also key.

References and Further Reading
- Digital Curation Centre http://www.dcc.ac.uk/
- Naming and Addressing: URIs, URLs, URNs http://www.w3.org/Addressing/
**Interoperability**

Interoperability is the ability of multiple systems with different hardware and software platforms, data structures, and interfaces to exchange data with minimal loss of content and functionality. Using defined metadata schemes, shared transfer protocols, and crosswalks between schemes, resources across the network can be searched more seamlessly. Describing a resource with metadata allows it to be understood by both humans and machines in ways that promote interoperability.

An important reason for choosing a standard metadata schema is to be able to interoperate with other collections. There are several different approaches to interoperability:

- **Cross-searching** - your metadata and images stay where they are, but are searched alongside other collections. This can be complicated to implement;

- **Contribution** - you give your metadata and images to someone who is building a larger collection. This is likely to require some simplification of any extended or adapted metadata into a more standard format;

- **Harvesting** - your metadata and images stay where they are, but you make available simple metadata records in a standard format for others to use in building catalogues which point to your resources.

The third model, harvesting, is a kind of cross between the other two. It has the advantage of opening up your collection to others without giving away control of your resources or necessarily requiring a great deal of effort. All JISC IE content providers are required to support either a distributed search interface or a metadata harvesting interface.

A number of initiatives\(^\text{100}\) are currently underway to specify, at a high level, how metadata standards for different domains (publications, individuals) should interoperate.

**Z39.50**

Z39.50 is a standard communications protocol for the search and retrieval of bibliographic data in online databases. Z39.50 is used on the Internet to search Copac (the UK Online Public Access Catalogue\(^\text{101}\)) of library holdings. The Zetoc\(^\text{102}\) electronic table of contents service Z39.50 client is based on the same technology as the Copac client.

The Z39.50 protocol is commonly used for cross-system search. Z39.50 implementers do not share metadata but map their own search capabilities to a common set of search attributes. A contrasting approach taken by the Open Archives Initiative is for all data providers to translate their native metadata to a common core set of elements and expose this for harvesting. A search service provider then gathers the metadata into a consistent central index to allow cross-repository searching regardless of the metadata formats used by participating repositories.

Z39.50 has been charged\(^\text{103}\) with being a heavyweight standard. There are lists of Z39.50 servers available, but the data is often incomplete, incorrect, or out of date. Even Z39.50 servers may be implemented in a non-standard way, or may have limited functionality. Therefore, the process of connecting to libraries using Z39.50 can be fraught, when the point of a standard should be to free the user from guesswork. In the UK, JISC services that implement Z39.50 include IESR, the Archives Hub, COPAC and Zetoc; the British Library, and many university libraries also have Z39.50 targets.

However, Z39.50 is not only used by libraries; in the cultural heritage sector the Consortium for the Computer Interchange of Museum Information (CIMI) worked to develop a Profile meeting the needs of cultural heritage practitioners\(^\text{104}\). The Government Information Locator Service (GILS) Profile\(^\text{105}\) also makes use of Z39.50 to link a wide range of resources internationally. Also, users connect their personal

---

\(^{100}\) e.g. [http://www.indecs.org](http://www.indecs.org), [http://www.cores-eu.net](http://www.cores-eu.net), and [http://www.w3.org/RDF](http://www.w3.org/RDF)

\(^{101}\) Copac library catalogue gives free access to the merged online catalogues of major University, Specialist, and National Libraries in the UK and Ireland: [http://copac.ac.uk/](http://copac.ac.uk/)

\(^{102}\) [http://zetoc.mimas.ac.uk/](http://zetoc.mimas.ac.uk/)

\(^{103}\) [http://www.librarything.com/](http://www.librarything.com/)

\(^{104}\) A Z39.50 Profile for Cultural Heritage Information [http://www.cimi.org/standards/index.html#THREE](http://www.cimi.org/standards/index.html#THREE)

\(^{105}\) Application Profile for the Government Information Locator Service (GILS), version 2 [http://www.gils.net/prof_v2.html](http://www.gils.net/prof_v2.html)
bibliographic reference managers, e.g. Endnote, to citation services such as Zetoc using Z39.50, although they are probably unaware of the Z39.50 protocol or the fact that they are using it.

**OpenURL**

The OpenURL\(^{106}\) standard enables the transfer of metadata about an item (a journal article or book, for example) from a resource, where a citation is discovered, to a link resolver. OpenURLs provide a means to tell another system what something is, rather than where it is located on the Internet (the function of a normal URL). In this way, via a link resolver, the OpenURL enables a user to obtain immediate access to the "most appropriate" copy of that object for them based on user and organisational preferences regarding the location of the copy, its cost, or agreements with information suppliers. OpenURL version 1.0 is NISO standard Z39.88 – 2004.

OpenURL linking not only improves the online working environment for users by reducing the number of linking dead ends but it also, by improving content visibility, could increase the use of licensed and subscribed materials and potentially reduce document delivery spend for libraries. OpenURL creates a persistent link, and can contain a DOI. It actually exists as a webpage (it is not a redirect) and can therefore be bookmarked easily and it can be generated from citation information without permissions. If it contains a DOI the an OpenURL resolver will be able to determine the details and hence location of the resource more precisely by using the DOI metadata.

Several different OpenURLs can point to the same paper or the same OpenURL can point to many different documents depending on the user’s context (i.e. how a resolver is set-up). This may make it difficult to connect information relating to the same publication by inspecting a bookmarked OpenURL. OpenURL should really only be used as a means to locate an appropriate copy and not as an identifier. Several JISC services act as OpenURL sources including Zetoc, Copac, Web of Knowledge and JSTOR.

**OAI-PMH**

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is the key standard being used to achieve interoperability through harvesting metadata.

OAI-PMH requires generation of XML records with metadata encoded as simple Dublin Core (although it is possible to include other schemas as well). These records are placed in a public space on a server where they are available for others to harvest (via the OAI protocol). They can then be incorporated into catalogues or directories. Although these OAI records will represent a simplified version of any richer metadata, users will have to link through to the original collection to view the digital resource. Once there, they will be exposed to the full metadata record and will be able to see the item in context. A practical example of this is the OAIster\(^{107}\) project at the University of Michigan. OAIster has harvested almost 10 million OAI-PMH records from 700 institutions.

Recently there has been a certain amount of scepticism expressed about the worth of the OAI-PMH protocol\(^{108}\). Amongst cons mentioned on mailing lists are that OAI-PMH is not suitable to transfer large-size files, especially when a network’s bandwidth is not large. OAI also has limitations when used for images and museum objects, which has implications for harvesting metadata, searching, and disclosure. An important outcome of the FAIR programme\(^{109}\) was a set of discussion papers exploring these issues. The conclusion to this work was that the use of collection description maybe more appropriate for some museum objects, as item descriptions maybe too similar to be useful.

The PerX project\(^{110}\), a JISC Digital Repositories Programme project, has been looking into subject resource discovery across a series of digital repositories. The project carried out a brief case study\(^{111}\) intended to illustrate some of the types of issues encountered by OAI-PMH service providers attempting to utilise third party metadata obtained via OAI-PMH.

---

107 [http://oaister.umdl.umich.edu/o/oaister/](http://oaister.umdl.umich.edu/o/oaister/)
108 [http://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=jisc-repositories](http://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=jisc-repositories)
109 [http://jisc.ac.uk/whatwedo/programmes/fair/synthesis/reptypes.aspx](http://jisc.ac.uk/whatwedo/programmes/fair/synthesis/reptypes.aspx)
110 [http://www.icbl.hw.ac.uk/perx/](http://www.icbl.hw.ac.uk/perx/)
111 [http://www.icbl.hw.ac.uk/perx/casestudyoxford.htm](http://www.icbl.hw.ac.uk/perx/casestudyoxford.htm)
As with METS, it may require quite a lot of work to generate the OAI records, although collection management software (e.g. digital repository systems) is incorporating OAI functionality. The JISC Technical Standards specify that service components must use OAI-PMH version 2.0 for metadata harvesting.

STARGATE\textsuperscript{112} (Static Repository Gateway and Toolkit), a JISC project examining the potential benefits of OAI-PMH Static Repositories as a means of enabling small publishers to participate more fully in the information environment, ran from 2005 - 2007. The project demonstrated the ease of use of OAI-PMH.

**HTTP (S)**

The Hypertext Transfer Protocol\textsuperscript{113} is the underlying communications protocol used by the Internet. HTTP defines how messages are formatted and transmitted, and what actions Web servers and browsers should take in response to various commands. For example, when you enter a URL in your browser, this actually sends an HTTP command to the Web server directing it to fetch and transmit the requested Web page. HTTP language is simple and understandable and is not difficult to follow. The current version (since June 1999) is HTTP/1.1.

HTTP is called a stateless protocol because each command is executed independently, without any knowledge of the commands that came before it. This is the main reason that it is difficult to implement Web sites that react intelligently to user input and retain information about user needs. This shortcoming of HTTP is being addressed in a number of technologies, including ActiveX, Java, JavaScript and cookies.

**AJAX**

AJAX, Asynchronous Javascript + XML\textsuperscript{114}, is a group of interrelated web techniques used for creating interactive web applications. AJAX is commonly used in Web2.0 applications to make web pages more dynamic. One of the frustrations for users of traditional HTML-based websites has been the time spent waiting for pages to reload and refresh after the user has chosen an option or clicked on a hyperlink. Several attempts have been made over the years to improve the dynamism of web pages through individual techniques such as use of Javascript, hidden frames, or Dynamic HTML (DHTML). With Ajax, only small amounts of information pass to and from the server once the page has first been loaded. This allows a portion of a webpage to be dynamically reloaded in real-time and creates the impression of richer, more 'natural' applications with the kind of responsive interfaces that are commonly found in desktop applications (Google calendar\textsuperscript{115} is an example of this).

**Service registries**

Registries are an important tool for managing metadata. Metadata registries can provide information on the definition, origin, source, and location of data. Registration can apply at many levels, including schemes, usage profiles, metadata elements, and code lists for element values. The metadata registry provides an integrating resource for legacy data, acts as a lookup tool for designers of new databases, and documents each data element.

Registries can also document multiple schemes or element sets, particularly within a specific field of interest.

**The Information Environment Service Registry**

The JISC Information Environment Service Registry\textsuperscript{116} (IESR) is a machine-to-machine middleware shared service providing a single central catalogue of descriptions of collections of resources available to researchers, learners and teachers in the UK, along with details of the services that provide access to those collections. The collections and services are described according to a set of metadata, which is defined by IESR, but is based on open standards wherever possible.

\textsuperscript{112}http://cdlr.strath.ac.uk/stargate/Index.htm
\textsuperscript{113}http://www.w3.org/Protocols/
\textsuperscript{114}For a detailed overview of Ajax and its application in Web 2.0 services see http://www.openajax.org/whitepaper.html
\textsuperscript{115}www.google.com/calendar
\textsuperscript{116}http://iesr.ac.uk/
The IESR has been developed with portals and other applications in mind, so that portal developers can access up-to-date information about available resources, without having to maintain this information themselves. A portal will be able to discover a collection of interest to an end-user, for example within a particular subject domain, and then provide to the end-user a link to the collection or a distributed search including it.

The aim is to supply enough technical information so that portal-builders and developers of other applications can use the IESR to provide (for example) cross-searching facilities over those resources that are of interest to the users of the portal/application. All significant collections and services made available within the JISC IE will be described according to the specifications being developed by the JISC IESR.

All JISC IE service components that require knowledge about available collections and services (e.g. portals, brokers and aggregators) have been asked to consider how they will make use of the JISC IESR. A variety of technical access methods can be described in the IESR, including OAI-PMH repositories, Z39.50 access to library OPACs and Web Services.

In a presentation to the JISC Repositories and Preservation Programme meeting in November 2007, Ann Apps, from IESR Research and Development at MIMAS, outlined the following benefits for using IESR:

- It provides a single place to find resources;
- It provides a single publication place;
- Resources come ‘online’ as they are registered;
- End users:
  - Only need to use the portal as a single place for discovery; and
  - May discover resources they were previously unaware of;
- Repositories and other Collections may benefit from increased use.

If machine-to-machine services, which use the various technologies described in this document, were registered with IESR then information about them could be more easily discovered and used.

IESR now has a web services interface (SRU/SRW). IESR currently holds over 5000 collection descriptions. These include JISC Collections, the resources identified by the JISC Digital Repositories and Archives Inventory (DRAI) and the details of operational repositories from OpenDOAR. Currently details of ESDS International Micro and Macro datasets are harvested weekly from the UK Data Archive, as are the details of repositories from OpenDOAR.

The Information Environment Metadata Schema Registry

The IEMSR project is funded by JISC through its Shared Services Programme. The IEMSR project is developing a metadata schema registry as a pilot shared service. The IEMSR has been developed as an application of the Resource Description Framework (RDF), a W3C recommendation (or set of recommendations) for a language to represent information about resources.

Metadata schema registries enable the publication, navigation and sharing of information about metadata. The IEMSR will act as the primary source for authoritative information about metadata schemas.

Johnston (2005) gave a useful description of a metadata schema registry. He states that a metadata schema registry is an application that provides services based on information about ‘metadata terms’ and about related resources. The term ‘metadata term’ is used here to denote some identified ‘unit of meaning’ deployed in a metadata description. Such terms are typically defined and managed as functional aggregations created to support some operation or service.

117 http://epub.mimas.ac.uk/papers/2007/jiscrrpm200711/apps-jiscrrpm-20071128.ppt#256_1
118 http://www.ukoln.ac.uk/projects/iemsr/
Johnston (2005) goes on to describe the services offered by a metadata schema registry. These may cover many different functions, and different metadata schema registries may provide different sets of functions depending on their purpose, scope and context; those functions might include:

1. Disclosure/discovery of information about metadata terms
2. Disclosure/discovery of relationships between metadata terms
3. Mapping or inferencing services based on relationships between terms
4. Verification of the provenance or status of metadata terms
5. Disclosure/discovery of related resources (such as functional aggregations of terms, guidelines for use, bindings for metadata instances, etc)

Metadata within the JISC IE is based largely on two key standards: the Dublin Core and the IEEE Learning Object Metadata (LOM\textsuperscript{120}) standard. The IEMSR will provide the JISC IE with a single point of referral for recommended schemas. It will allow various initiatives within the JISC IE to publish "application profiles"\textsuperscript{121} of these standards in a common registry, making them available to others. This provides a concrete way of encouraging sensible uniformity alongside necessary divergence. It helps avoid unnecessary duplication of effort, and supports sharing of common approaches.

OpenDOAR

OpenDOAR\textsuperscript{122} is a directory of academic open access OAI-PMH-based repositories. OpenDOAR maintains a comprehensive and authoritative list of institutional and subject-based repositories. It also encompasses archives set up by funding agencies like the National Institutes for Health in the USA or the Wellcome Trust in the UK and Europe. A directory such as OpenDOAR makes it easier to identify and mine individual repositories. Based on the results of a review of twenty four similar directories\textsuperscript{123}, OpenDOAR emerged as the clear leader in terms of scope and usability.

OpenDOAR forms a major target resource for JISC services such as Intute Repository Search\textsuperscript{124} and the Depot\textsuperscript{125}. By November 2007 there were 1000 entries from across the globe, making OpenDOAR a significant step forward in enabling the global virtual repository network to cooperate in new and innovative ways\textsuperscript{126}.

Web services

WSDL

Web Services Description Language\textsuperscript{127} (WSDL) is an XML grammar that defines the functionality offered by a Web service and the format of messages sent and received by the Web service. A Web service's WSDL document defines what services are available in the Web service. The WSDL document also defines the methods, parameter names, parameter data types, and return data types for the Web service. An application that uses a Web service relies on the Web service's WSDL document to access the Web service's features.

\begin{footnotesize}
\begin{enumerate}
\item http://ieeeltsc.org/
\item http://www.ariadne.ac.uk/issue25/app-profiles/
\item http://www.opendoar.org/
\item http://www.intute.ac.uk/irs/
\item http://depot.edina.ac.uk/
\item http://www.w3.org/TR/wsdl
\end{enumerate}
\end{footnotesize}
**UDDI**
Universal Description, Discovery, and Integration is an XML-based registry for businesses worldwide to list themselves on the Internet. Its ultimate goal is to streamline online transactions by enabling companies to find one another on the Web and make their systems interoperable for e-commerce. The UDDI specification offers a framework for Web services integration and has also adopted early versions of the proposed Simple Object Access Protocol (SOAP) messaging guidelines for cross platform programming.

**SOAP**
SOAP (Simple Object Access Protocol) is a way for a programme running in one kind of operating system (Windows) to communicate with a programme in the same or another kind of an operating system (such as Linux) by using HTTP and XML as the mechanisms for information exchange. Since Web protocols are installed and available for use by all major operating systems, HTTP and XML provide a solution to the problem of how programmes running under different operating systems in a network can communicate with each other. SOAP specifies exactly how to encode an HTTP header and an XML file so that a program in one computer can call a program in another computer and pass it information. It also specifies how the called program can return a response. SOAP has been proposed as a standard interface to the Internet Engineering Task Force (IETF).

**REST**
Representational State Transfer (REST) is an approach for getting information content from a website by reading a designated webpage that contains an XML file that describes and includes the desired content. For example, REST could be used by an online publisher to make syndicated content available. Periodically, the publisher prepares and activates a webpage that includes content and XML statements that describe the content. Subscribers would need only to know the URL for the page where the XML file was located, read it with a Web browser, interpret the content data using the XML information, and reformat and use it.

REST exploits the existing technology and protocols of the Web. REST is simpler to use than the SOAP approach, which requires writing or using a provided server program (to serve data) and a client program (to request data). SOAP, however, offers potentially more capability.

**References and Further Reading**
- OAI-PMH 2.0, [http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm](http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm)
- A list of registered OAI data providers is available at [http://www.openarchives.org/Register/BrowseSites](http://www.openarchives.org/Register/BrowseSites)
- OAIlster is an example of an OAI harvesting service which provides aggregation and search, [http://oaister.umdl.umich.edu/o/oaister/](http://oaister.umdl.umich.edu/o/oaister/)
- One of the Key objectives of JISC CETIS is to “establish interoperability specifications, standards and application profiles that meet the needs of the JISC community”. Recent discussions can be accessed on the CETIS wiki: [http://wiki.cetis.ac.uk/CETIS_Wiki](http://wiki.cetis.ac.uk/CETIS_Wiki)
Resource Discovery

Resource Discovery is taken to mean the process of identifying and accessing information relevant to the searcher, although identification (discovery) and access are best considered as separate processes.

“Finding relevant information on the World Wide Web has become increasingly problematic due to the explosive growth of networked resources. Current Web indexing evolved rapidly to fill the demand for resource discovery tools, but that indexing, while useful, is a poor substitute for richer varieties of resource description.” Dublin Core metadata for resource discovery, September 1998

Effective cataloguing and metatagging are required for:

- allowing resources to be found by relevant criteria;
- identifying resources;
- bringing similar resources together;
- distinguishing dissimilar resources; and
- giving location information (such as a URL).

Machine-based resource discovery relies on crawling, clustering, and classifying resources discovered on the Web automatically. Resources are organized with respect to metadata that characterize their content (for data sources), their semantics (in terms of ontological classes and relationships), their characteristics (syntactical properties), their performance (with metrics and benchmarks), their quality (curation, reliability, trust), etc. Resource discovery systems allow the expression of queries to identify and locate resources that implement specific tasks.

There are currently two main mechanisms for enabling discovery of content, but both are facilitated via a service element. The two approaches are harvesting and federation:

- Harvesting involves the aggregation of metadata from dispersed repositories via syndication standards, and then providing a single service to search the pool of resultant metadata, followed by direct access to the content. The standards and specifications for harvesting are OAI-PMH, ATOM and RSS;
- Federation is the aggregation of multiple search services, where the same search terms are sent to each individual service and the result metadata is returned via a single service, followed by direct access to the content. The standards and specifications for federated search are Z39.50, SRU and SRW.

JISC Collections negotiate licenses for access to content held at various organisations. Each organisation has to abide by a set of technical guidelines. At the moment only SRU/ SRW are supported within their technical specifications (federation approach). Some organisations may wish them to incorporate syndication standards and specifications such as OAI-PMH, ATOM and RSS (for harvesting) in future; JISC Collections will be responsive to user needs in this respect.

What is of the greatest importance though is for all content to be readily accessible - accessible in terms of users being able to access it seamlessly, whether it originated locally, was provided by another institution, was content licensed to the institution or whatever other scenario there might be. Metadata should be provided for each piece of content.

Search services should be written so that they are compatible with the OpenSearch specification\(^\text{128}\) for plugging into the latest generation of web browsers. OpenSearch is a collection of simple formats for the sharing of search results. The OpenSearch description document format can be used to describe a search engine so that it can be used by search applications. OpenSearch response elements can be used to extend existing syndication formats, such as RSS and Atom, with the extra metadata needed to return search results.

\(^{128}\) [http://www.opensearch.org/Home](http://www.opensearch.org/Home)
Finally, if content is to be made available more widely than the local system through harvesting or federation approaches, the IPR and Copyright of the content items should be upheld at all times.

**Distributed searching**

Where distributed searching is used within the JISC IE, JISC Technical Guidelines state that service components must use either Z39.50 version 3 or SRW/SRU version 1.2.

**Z39.50 and Z39.50 Bath Profile**

Z39.50 is a mature international standard that specifies searching and retrieving information from remote databases. For further details on Z39.50, see section on Interoperability.

Within any standard there are implementation options; and the meaning of some specifications may be open to interpretation. Software developers have selected different options to implement or interpreted the standard differently in their systems. This may result in the user receiving many false hits or, conversely, not retrieving a record even if it is in the database. Profiles provide a mechanism to specify a standard way to interpret and implement options within the standard. The Bath Profile is an internationally registered Z39.50 profile supporting library applications and resource discovery. The Profile defines searching across multiple servers to improve search and retrieval among library catalogues, union catalogues, and other electronic resources worldwide.

**SRU and SRW**

SRU\(^{129}\) is a web protocol based on Web 2.0 technology that uses elements of Z39.50 with results returned as XML. SRW is a further development, an open protocol that uses web service standards for the communication of metadata and data.

SRU stands for Search/Retrieve via URL and does not use SOAP; it is currently in version 1.2, which also subsumes SRW. SRW stands for Search/Retrieve Web Service and is the sister approach to SRU, which uses standardised SOAP calls to return the same results. It is now referred to as a SOAP binding of SRU.

SRW/U and OAI-PMH are complementary protocols. They have similar goals, namely, the retrieval of metadata from remote hosts, but each provides functionality that the other does not. Both protocols have similar ‘about’ functions. SRW/U’s explain operation and OAI-PMH’s identify verb both return characteristics describing the properties of the remote service. Both protocols have a sort of “browse” functionality. SRW/U has its scan function and OAI-PMH has ListSets. Scan is like browsing a book’s back-of-the-book index. ListSets is similar to reading a book’s table of contents.

SRW/U and OAI differ when it comes to retrieval. SRW/U provides a much more granular approach (precision) at the expense of constructing complex Common Query Language (CQL) queries. OAI-PMH is stronger on recall allowing a person to harvest the sum total of data a repository has to offer using a combination of the ListRecords and GetRecords verbs. This is implemented at the expense of gathering unwanted information.

If a set of data were exposed via SRW/U as well as OAI-PMH, then SRW/U would be the tool to use if a person wanted to extract only data crossing predefined sets. OAI-PMH would be more relevant if the person wanted to get everything or predefined subsets of the data.

**OAI-ORE**

Although metadata within Digital Repositories can be effectively searched and harvested through techniques such as OAI-PMH, SRU/SRW etc., there are no current widely adopted standards for exchanging/transferring the actual digital objects (datastreams) attached to the metadata (images/video/sound etc). This is however being addressed, and foremost amongst those addressing this

---

\(^{129}\) [http://www.loc.gov/standards/sru/](http://www.loc.gov/standards/sru/)
issue is Herbert Van de Sompel working at the Los Alamos National Library in the USA on the OAI-ORE\textsuperscript{130} project. Other interested parties include Microsoft, the Andrew Mellon Foundation, the Coalition for Networked Information, the Digital Library Federation and JISC.

**Digital repositories**

One of the major objectives is to develop standards that can be used for the interoperability of digital objects between digital repositories. The current problem is that across the digital community, there is a different perspective on the design and management of digital repositories. Repository systems use different software platforms, serve different communities and have different policies for deposit, storage, preservation etc resulting in a lack of cross-repository interoperability.

A lot of JISC activity has been funded in this area; the Digital Repositories Programme from 2005-2007; and the Repositories and Preservation Programme from 2006-2009. As part of the Digital Repositories Programme, JISC established the Repositories Research Team to provide support to technical development staff within the Programme\textsuperscript{131}. The remit for the work of the research team is quite wide and includes helping projects find and exploit synergies across the programme and beyond, gathering scenarios and use cases from projects, liaising with other national and international repositories activities, including liaison with the e-Framework, synthesizing project and programme outcomes, and engaging with interoperability standards activity and repository architectures. Since late 2006 the JISC Repository Support Project\textsuperscript{132} has supported repository managers and officers, and offers introductory guidance on standards in this area.

**Semantic Web Services**

**RDF**

Resource Description Framework\textsuperscript{133} (RDF) is a specialist XML application (created in 1999) which provides a generic framework for describing metadata and therefore encoding knowledge. RDF was updated in 2004. RDF is no longer only used for encoding information about web resources, but also information about and relations between things in the real world: people, places, concepts, etc. RDF provides a general, flexible method to break up knowledge into small pieces, called triples, with some rules about the meaning of those pieces.

Triples have the form:

\[
\text{subject} \ {\text{predicate}} \ {\text{object}}
\]

i.e. if \{Document X\} \{has author\} \{Charles Dickens\} and \{Document X\} \{has title\} \{Little Dorrit\}; then a computer can use logic to assert that \{Little Dorrit\} \{has author\} \{Charles Dickens\}

In the Semantic Web, computers do the browsing (and searching, and querying) for us. The Semantic Web enables computers to seek out knowledge distributed throughout the Web, mesh it, and then take action based on it. RDF is thought to be a key standard in the development of the Semantic Web\textsuperscript{134}.

\textsuperscript{130} OAI-ORE Project http://www.openarchives.org/ore/
\textsuperscript{131} http://www.ukoln.ac.uk/repositories/digirep/index/About#Repositories_Reasearch_Team
\textsuperscript{132} http://www.rsp.ac.uk
\textsuperscript{133} http://www.w3.org/RDF/
\textsuperscript{134} http://www.rdfabout.com/intro/?section=contents
Emergent Standards

**RDFS**
RDF Schema is an extensible knowledge representation language which introduces the notion of a class. A class is a type of thing, i.e. computers are members of the class Machine.

**OWL**
Web Ontology Language (OWL) is also a knowledge representation language. OWL defines more classes that let RDF authors define more of the meaning of their predicates within RDF. OWL is considered one of the fundamental technologies underpinning the Semantic Web, and has attracted both academic and commercial interest.

Common Vocabularies

**FOAF**
“Friend-of-a-friend” is a machine-readable ontology describing people, their activities and their relations to other people and objects. Anyone can use FOAF to describe him or herself. FOAF allows groups of people to describe social networks without the need for a centralised database. FOAF is an extension to RDF and is defined using OWL.

**SKOS**
Simple Knowledge Organisation Systems aim to allow expression of the basic structure of controlled vocabularies via a standardised notation for the attribution of classes of resource and properties of those classes.

Syndication Formats

The web has become enormous, and search engines that crawl the surface of the web are picking up only a small fraction of the available content. Further, some of the richest and most interesting content cannot be crawled and indexed by one search engine or navigated by one relevancy algorithm\(^\text{135}\). RSS and ATOM are lightweight options for syndicating metadata compared to other methods of exposing or exchanging metadata, such as OAI-PMH, or returning the results of various queries. Both RSS and ATOM can also be used for providing updates from institutional and open access repositories to the user.

**RSS**
RSS (Rich Site Summary or Real Simple Syndication\(^\text{136}\)) is a format for delivering regularly changing web content. Many news-related sites, weblogs and other online publishers syndicate their content as an RSS Feed to whoever wants it. RSS solves a problem for people who regularly use the web. It allows a user to stay informed by retrieving the latest content from the sites they are interested in, and so users save time by not needing to visit each site individually. Privacy is ensured by not needing to join each website’s email newsletter.

An RSS news reader\(^\text{137}\) is required to check RSS web feeds and for the user to read any new articles that have been added to them. There are two types of news readers, those that are downloaded locally, and those that are hosted on the web and accessed through a browser.

---

\(^\text{135}\) http://www.opensearch.org/

\(^\text{136}\) http://www.w3schools.com/rss/rss_intro.asp

\(^\text{137}\) For a list of RSS readers see: http://www.jisc.ac.uk/rss_feeds/web_feed.aspx
**ATOM**

ATOM\textsuperscript{138} is also written in XML and was created as a response to perceived deficiencies of RSS 2.0 and the ‘version wars’ in the RSS community\textsuperscript{139}. It uses unique identifiers and is validated by a schema.

**OpenSearch**

Different types of content require different types of search engines. The best search engine for a particular type of content is likely to be the search engine written by the people that know the content the best. OpenSearch helps search engines and search clients communicate by introducing a common set of formats to perform search requests and syndicate search results. OpenSearch was created by A9.com, an Amazon.com company, and the OpenSearch format is now in use by hundreds of search engines and search applications around the Internet. The OpenSearch specification is made available according to the terms of a Creative Commons license.

**Portals and port-lets**

Portlet specifications can be used to supplement the delivery of existing services through emerging technologies. The CREE project undertook an investigation of the JSR168 and WSRP standards as part of its technical evaluation\textsuperscript{140}. Both of these standards are aimed at portalising existing services and addressing the presentation of information in different environments. In analysing the potential use of JSR 168 and WSRP, it is necessary to have a portal framework within which portlets described using these two standards can be surfaced and tested. The portal framework must be compliant with one or both of the two standards itself.

**JSR 168**

JSR 168\textsuperscript{141} enables interoperability between portlets and portals and is used for embedding information in institutional portals. JSR 168 is a Java Community Process standard that specifies a set of APIs to enable components of a portal, or portlets, to be developed. The standard seeks to enable these portlets, once developed, to then be used within any JSR 168 compliant portlet container. Many available portlet containers currently available and used within portals are Java based, so wide interoperability and use is possible.

**WSRP**

Web Services for Remote Portlets\textsuperscript{142} (WSRP) is the OASIS standard Web Services for Remote Portlets. This web services standard seeks to enable ‘plug-n-play’ portals, facilitating the surfacing of access to search interfaces and many other areas of functionality within a portal. Unlike JSR 168 it is platform-independent, and services to be surfaced can originate from many different systems and environments. WSRP uses SOAP as a means to communicate between a WSRP producer and a WSRP consumer.

**Presentation layers**

Many presentation standards fulfil the criteria of default standards that need to, or should, be used for delivery via the Web (HTML, XHTML in some cases).

Presentation is also key to accessibility and much presentation work has been based on the need to address accessibility issues\textsuperscript{143}. The use of accessibility standards is an evolving one, guided by the latest guidelines available. WAI-WCAG\textsuperscript{144} is widely followed, though is clearly stated as current best practice only, not

---

\textsuperscript{138} http://www.atomenabled.org/
\textsuperscript{139} A full comparison of the technical differences between ATOM and RSS can be viewed at http://www.intertwingly.net/wiki/pie/Rss20AndAtom10Compared.
\textsuperscript{140} See use cases at: http://www.hull.ac.uk/cree/downloads/CREEtechnicalreportC.pdf
\textsuperscript{142} http://www.oasisopen.org/committees/tc_home.php?wg_abbrev=wsrp
\textsuperscript{143} http://www.w3.org/WAI/quicktips/
\textsuperscript{144} http://www.w3.org/WAI/
definitive. The adoption of accessibility standards is driven by external factors, most often legal factors, though there remains ongoing debate about the best ways to achieve the accessibility required.

Location and delivery
It is worth noting that resource discovery is usually only the first stage of the process: the user will then require access to the digital object (if not the physical object as well) and may wish to make a copy to store local to them for further re-use.

References and Further Reading
Conclusions for Use of Standards

There will always be a pool of standards for projects to choose from but the important thing is for creators to be strictly consistent in their application - your original may be superseded but then you have a better chance of migrating without data loss. Additionally once you can map a standard you can create crosswalks and gateways for interoperability.

It is worth noting that many standards that deal with interoperability could also be termed communication standards. As the eReSS project\[145\] states: “…[communication] standards …facilitate many different levels of communication: email communication (IMAP, POP, SMTP), videoconferencing (H.323, RTP, RTSP), internal software communication (JMS, SOAP, WS-*), web communication (HTTP(S)), communication of search parameters (Z39.50, SRW/U), communication of metadata (OAI-PMH, RSS), standards that facilitate communication through identification (URI), and file transfer (WebDAV). Many of these standards also determine the organisation of information and metadata.

Recommendations for Use of Standards

The use of standards is, of course, encouraged, but mandating the use of the same systems or the same vocabularies is neither possible nor desirable. However, the wider adoption of improved best practices should greatly facilitate the incremental development of interoperability.

\[145\] http://www.hull.ac.uk/esig/eress.html
# Glossary of terms

This Glossary does not attempt to be all encompassing, but to provide a guide to terms used with in the SCA Standards Guide. Readers and users should also be able to search for terms online (“Google” them), and cross reference terms used against the most up-to-date resources found on the Internet. JISC Advisory Services also maintain useful glossaries of terms which can be found through their websites.

Users of the SCA Standards Guide and this glossary could also consult the following resource: Glossary of Internet Terms: [http://www.matisse.net/files/glossary.html](http://www.matisse.net/files/glossary.html); this is regularly updated.

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access</td>
<td>The activity of obtaining the most appropriate copy of a discovered resource.</td>
</tr>
<tr>
<td>Aggregator</td>
<td>A structured network service that gathers metadata from a range of other, heterogeneous, local or remote structured network services. Aggregators are intended for use by software applications. In the context of the JISC IE, aggregators interact with indexes, catalogues, content providers and other aggregators using the OAI-PMH and RSS/HTTP. Aggregators interact with portals using the OAI-PMH. In some cases an aggregator may offer its aggregated metadata as a Z39.50 target.</td>
</tr>
<tr>
<td>Alert</td>
<td>A disclosure strategy based on issuing information about new or updated resources, typically through a network service or via email.</td>
</tr>
<tr>
<td>API</td>
<td>Application Programming Interface. A library of functions that are pre-written (possibly by a third party) to offer functionality to the programmer.</td>
</tr>
<tr>
<td>Application profile</td>
<td>A set of elements with associated descriptions of usage in a particular context.</td>
</tr>
<tr>
<td>Athens</td>
<td>Athens is an Identity Management service that is run by Eduserv and funded by JISC. Athens is currently being replaced as the national access management service for the UK educational sector by a new service based on the Shibboleth technology. JISC will stop funding Athens in 2008. Athens will still be available after 2008 but as a subscription service. For more information about Athens: <a href="http://www.athensams.net/">http://www.athensams.net/</a></td>
</tr>
<tr>
<td>AthensDA</td>
<td>Athens Devolved Authentication. An Identity Management service, run by Eduserv, that enables an organisation to maintain a single set of credentials for a user, and for that user to be authorised for access to online services depending on permissions defined by the user’s organisation. For more information about AthensDA: <a href="http://www.athensams.net/developmen%20t/devolved_authentication/">http://www.athensams.net/developmen%20t/devolved_authentication/</a></td>
</tr>
<tr>
<td>Attribute</td>
<td>Information about an individual in defined formats such as member of organisation x, member of department y, role equals student or faculty.</td>
</tr>
<tr>
<td>Authentication Service</td>
<td>A structured network service that determines that the digital ID being presented to a network service is being used by the real-world individual who has the rights to use it. This is often achieved through the use of a username/password combination or a digital certificate, depending on the degree of assurance required.</td>
</tr>
<tr>
<td>Authentication</td>
<td>The process of verifying who is requesting access to a resource.</td>
</tr>
<tr>
<td>Authorisation Service</td>
<td>A structured network service that indicates whether a particular digital ID has the necessary access-rights to access a particular resource.</td>
</tr>
<tr>
<td>Authorisation</td>
<td>The process of determining whether access should be granted to an individual based on information about that individual.</td>
</tr>
<tr>
<td>BECTa</td>
<td>British Educational Communications and Technology Agency (formerly NCET). BECTa co-ordinates educational and technological developments across the compulsory and post-compulsory educational communities. BECTa also seeks to evaluate information and communications technology (ICT) practice; support existing applications of ICT; and investigate emerging technologies and associated pedagogy</td>
</tr>
<tr>
<td>Broker</td>
<td>A structured network service that provides (search) access to a range of other, heterogeneous, local or remote structured network services. Brokers are intended for use by software applications. In the context of the JISC IE, brokers interact with indexes, catalogues, aggregators, content providers, other brokers and portals using</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>-------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Z39.50</td>
<td>A survey or discover strategy based on following links between resources.</td>
</tr>
<tr>
<td>BSI</td>
<td>British Standards Institute</td>
</tr>
<tr>
<td>Catalogue</td>
<td>A network service that provides access to a collection of human-generated metadata records.</td>
</tr>
<tr>
<td>CETIS</td>
<td>The Centre for Educational Technology Interoperability Standards – formerly known as the UK IMS project</td>
</tr>
<tr>
<td>Citation</td>
<td>A quoting of an authoritative source for substantiation</td>
</tr>
<tr>
<td>CLA</td>
<td>Copyright Licensing Agency</td>
</tr>
<tr>
<td>Collection</td>
<td>An aggregation of one or more items.</td>
</tr>
<tr>
<td>Compliance</td>
<td>The ability to operate in a way defined by a standard</td>
</tr>
<tr>
<td>Conformance</td>
<td>The ability to operate in a way defined by a specification</td>
</tr>
<tr>
<td>Content Provider</td>
<td>A network service that makes a collection available. A content provider may disclose metadata about its resources through a structured network service. In the context of the JISC IE, a content provider interacts with brokers, aggregators and portals using Z39.50, the OAI-PMH and RSS/HTTP. Note that 'content provider' may also refer to the organisation that makes collections available - which may be a JISC-funded service, an HE/FE institution, or some other organisation.</td>
</tr>
<tr>
<td>Controlled vocabulary</td>
<td>A collection of controlled terms; a formally maintained list of terms intended to provide values for metadata elements</td>
</tr>
<tr>
<td>Copyright</td>
<td>One of the suite of Intellectual Property Rights</td>
</tr>
<tr>
<td>Cross walk</td>
<td>A mapping from the elements of one namespace to the elements of another namespace.</td>
</tr>
<tr>
<td>Data</td>
<td>Information that is recorded or processed</td>
</tr>
<tr>
<td>Data Controller</td>
<td>Individual or organisation with final control to gather and process data</td>
</tr>
<tr>
<td>Data Mining</td>
<td>The seeking and gathering of individual pieces of data from different sources, to compile a more detailed profile of a data subject</td>
</tr>
<tr>
<td>DCMES</td>
<td>Dublin Core Metadata Element Set</td>
</tr>
<tr>
<td>DCMI</td>
<td>Dublin Core Metadata Initiative</td>
</tr>
<tr>
<td>Deposit</td>
<td>The activity of placing a resource within a repository, typically with the intention of making it available to others.</td>
</tr>
<tr>
<td>Disclosure</td>
<td>The function of offering metadata records for searching (using Z39.50) or gathering (using the OAI-PMH and/or RSS/HTTP) through a structured network service.</td>
</tr>
<tr>
<td>Discover</td>
<td>The activity of finding items and/or services of interest.</td>
</tr>
<tr>
<td>DOM</td>
<td>Document Object Model</td>
</tr>
<tr>
<td>Dublin Core</td>
<td>Shorthand term for the Dublin metadata core element set, which is a core list of metadata elements agreed at the OCLC / NCSA Metadata Workshop in Dublin, Ohio in March 1995. The Dublin Core is positioned as a simple information resource description supporting resource discovery. It also aims to provide a basis for semantic interoperability between other, more complicated formats.</td>
</tr>
<tr>
<td>Early Adopters</td>
<td>Institutions who become early adopters (opens in new window) of the next generation of access management tools.</td>
</tr>
<tr>
<td>EDINA</td>
<td>A JISC-supported 'datacentre'. They provide online services to the UK further and higher education communities. Edina built and maintained the UK pilot federation. Further details: <a href="http://edina.ac.uk/">http://edina.ac.uk/</a></td>
</tr>
<tr>
<td>Educause</td>
<td>Non-profit association in the US that promotes the use of information technology in higher education. Further details: <a href="http://www.educause.edu/">http://www.educause.edu/</a></td>
</tr>
<tr>
<td>Element</td>
<td>a formally defined attribute or category of description in a metadata set.</td>
</tr>
<tr>
<td>---------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Endnote</td>
<td>Software tool for publishing and managing bibliographies</td>
</tr>
<tr>
<td>Federated authentication</td>
<td>See Devolved Authentication, and Federation.</td>
</tr>
<tr>
<td>Federated search</td>
<td>A search agent that enables the user to search multiple sources at once and receive the search results on one list</td>
</tr>
<tr>
<td>Federation</td>
<td>A group or set of organisations that share a common set of policies and rules in order to establish common trust and language/terminology to aid cross-domain authentication and authorisation. The UK Federation is called The UK Federation for Education &amp; Research. It is run by JANET(UK) on behalf of JISC and Becta. For more information: <a href="http://www.ukfederation.org.uk/">http://www.ukfederation.org.uk/</a>.</td>
</tr>
<tr>
<td>FTP</td>
<td>File Transfer Protocol. FTP allows a person to transfer files between two computers, generally connected via the internet</td>
</tr>
<tr>
<td>Handle Service (HS)</td>
<td>The Shibboleth component that authenticates the user. It issues the Attribute Query Handle that is used later in the authorisation process to request user attributes. When a user is successfully authenticated, the HS presents the handle to the SHIRE in the form of a signed SAMLResponse, sent via an HTTP-POST.</td>
</tr>
<tr>
<td>Harvesting</td>
<td>Automated gathering of data from external sources.</td>
</tr>
<tr>
<td>HTML</td>
<td>Hyper-text Markup Language, used to mark up documents for use on the World Wide Web. It enables hypertext links to be followed, thus linking documents over the web.</td>
</tr>
<tr>
<td>Identity Provider</td>
<td>(IdP) In the Shibboleth architecture, the Identity Provider is the organisation that provides authentication for a user. Authorisation is provided by the Service Provider. Formerly known as the origin.</td>
</tr>
<tr>
<td>Index</td>
<td>A network service that provides access to a machine-generated database of information derived from the content of items in a collection.</td>
</tr>
<tr>
<td>Internet2</td>
<td>Provides a central resource to develop and deploy advanced network applications, and technologies for research and higher education. Internet2 is funded by 200 US universities. Further details: <a href="http://www.internet2.edu">http://www.internet2.edu</a></td>
</tr>
<tr>
<td>Interoperability</td>
<td>The ability of systems and data to work seamlessly together.</td>
</tr>
<tr>
<td>IP</td>
<td>Internet Protocol</td>
</tr>
<tr>
<td>ISO</td>
<td>International Standards Organisation</td>
</tr>
<tr>
<td>Item</td>
<td>A physical or digital entity.</td>
</tr>
<tr>
<td>JANET</td>
<td>Joint Academic NETwork. A private, government-funded network for education and research. All further and higher education organisations are connected to JANET, as are all the Research Councils.</td>
</tr>
<tr>
<td>JANET(UK)</td>
<td>Manages the operation and development of the JANET network. Runs the UK Federation for Education &amp; Research on behalf of JISC and Becta.</td>
</tr>
<tr>
<td>JISC</td>
<td>Joint Information Systems Committee. Supports further and higher education in the UK in the use of information and communications technology. Further details: <a href="http://www.jisc.ac.uk">http://www.jisc.ac.uk</a></td>
</tr>
<tr>
<td>Liberty Alliance</td>
<td>A project, formed in September 2001, to establish an open standard for federated network identity. This will be accomplished by developing technical specifications that support a broad range of identity-based products and network devices. It is a consortium of more than 150 technology and consumer organizations. Further details: <a href="http://www.projectliberty.org/index.php">http://www.projectliberty.org/index.php</a></td>
</tr>
<tr>
<td>Licence Grants</td>
<td>the power or authority to perform an act, which otherwise would be unlawful e.g. Licences can be granted by patent owners, to another party, allowing them to use the patented invention</td>
</tr>
<tr>
<td><strong>Metadata</strong></td>
<td>Information describing an information resource, sometimes referred to as data about data</td>
</tr>
<tr>
<td><strong>Metadata architecture</strong></td>
<td>&quot;a coherent collection of enabling technologies, element sets, and standards of practice that collectively support the creation, management and exchange of interoperable metadata.&quot;</td>
</tr>
<tr>
<td><strong>Metadata element</strong></td>
<td>An individual part of a metadata structure.</td>
</tr>
<tr>
<td><strong>Metadata record</strong></td>
<td>Structured data about a resource.</td>
</tr>
<tr>
<td><strong>Metadata registry</strong></td>
<td>A place to access metadata schema information.</td>
</tr>
<tr>
<td><strong>Metadata schema</strong></td>
<td>A description of the structure given to a set of metadata elements.</td>
</tr>
<tr>
<td><strong>Metadata Schema Registry</strong></td>
<td>A network service that stores and makes available information about the metadata schemas in use by other services.</td>
</tr>
<tr>
<td><strong>Middleware</strong></td>
<td>Network-based services that sit between users and the service that they are trying to access, enabling them to access that service or provide additional functionality. Authentication/authorisation is a classic example.</td>
</tr>
<tr>
<td><strong>Namespace</strong></td>
<td>A scoping device used for uniquely identifying registered entities.</td>
</tr>
<tr>
<td><strong>OAI-PMH</strong></td>
<td>The Open Archives Initiative Protocol for Metadata Harvesting. A protocol supporting the bulk exchange of metadata records between structured network services.</td>
</tr>
<tr>
<td><strong>OASIS</strong></td>
<td>Organization for the Advancement of Structured Information Systems. A standards body. A not-for-profit global consortium that drives the development, convergence and adoption of e-business standards.</td>
</tr>
<tr>
<td><strong>OpenSAML</strong></td>
<td>An open-source library implementing the SAML protocol. The project is currently hosted and controlled by Internet2.</td>
</tr>
<tr>
<td><strong>OpenURL</strong></td>
<td>A standard for transporting metadata between network services by encoding it in a URL.</td>
</tr>
<tr>
<td><strong>OpenURL Link Servers</strong></td>
<td>A network service that takes metadata about a resource in the form of an OpenURL and supplies links (pointers) to services on, or related to, that resource, for example delivery services.</td>
</tr>
<tr>
<td><strong>PERMIS</strong></td>
<td>Privilege and Role Management Infrastructure Standards Validation. A tool for determining the rights of a user to access a service through the analysis of user attributes. For more information: <a href="http://www.permis.org/">http://www.permis.org/</a>.</td>
</tr>
<tr>
<td><strong>Portal</strong></td>
<td>A network service that provides a personalised, single point of access to a range of heterogeneous network services, local and remote, structured and unstructured. Portal functionality often includes resource discovery, email access and online discussion fora. Portals are intended for (human) end-users using common Web 'standards' such as HTTP, HTML, Java and JavaScript. In the context of the JISC IE, portals interact with brokers, aggregators, indexes, catalogues and content providers using Z39.50, SRW, the OAI-PMH and RSS/HTTP.</td>
</tr>
<tr>
<td><strong>Provision / Content Provision</strong></td>
<td>The function of making a collection or item available through a network service.</td>
</tr>
<tr>
<td><strong>Publish</strong></td>
<td>The activity of making a resource available to others.</td>
</tr>
<tr>
<td><strong>PURL</strong></td>
<td>Persistent Uniform Resource Locator</td>
</tr>
<tr>
<td><strong>RDF</strong></td>
<td>Resource Description Framework</td>
</tr>
<tr>
<td><strong>Regional Support Centres (RSC)</strong></td>
<td>Advise on and promote the use of network learning technologies and resources in the UK tertiary education sector. Funded by JISC.</td>
</tr>
<tr>
<td><strong>Resource</strong></td>
<td>An item, collection or service of interest to the end-user.</td>
</tr>
<tr>
<td><strong>RSS/HTTP</strong></td>
<td>RSS is RDF Site Summary - an RDF/XML application for describing syndicated news feeds. RSS/HTTP is the exchange of RSS files between network services using the Hypertext Transfer Protocol.</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>SAML</td>
<td>Security Assertion Markup Language. A standard defined and maintained by OASIS. It's an XML-based framework for creating and exchanging security information between online parties.</td>
</tr>
<tr>
<td>SCA</td>
<td>The Strategic Content Alliance works with public sector bodies across the UK to develop a framework to share and access information.</td>
</tr>
<tr>
<td>Schema</td>
<td>Sets of statements expressed in a data definition language; A set of metadata elements representing the attributes of a resource</td>
</tr>
<tr>
<td>Search</td>
<td>A survey or discover strategy based on issuing a query and receiving a set of search results.</td>
</tr>
<tr>
<td>Semantic web</td>
<td>The Semantic Web is a concept of the Web where machines are able to understand and interpret web resources for people.</td>
</tr>
<tr>
<td>Semantics</td>
<td>The meanings of words</td>
</tr>
<tr>
<td>Service Provider (SP)</td>
<td>In the Shibboleth architecture, the Service Provider is the provider of information or resources. Formerly known as the target.</td>
</tr>
<tr>
<td>Service Registry</td>
<td>A network service that stores and makes available descriptions of (i.e. metadata about) services and the content of collections made available through those services. A service registry is used by portals to determine what collections are available to end-users, and by portals, brokers and aggregators to determine how to interact with available network services.</td>
</tr>
<tr>
<td>Service</td>
<td>The provision of, or system of supplying, one or more functions of interest to an end-user or software application. 'Informational' services provide access to, or metadata about, items and/or collections. 'Transactional' services are those that do not primarily concern the supply of information, for example photocopying, printing or banking services. Services may be physical or on-line (digital).</td>
</tr>
<tr>
<td>Shibboleth</td>
<td>An Internet2 project to define an architecture that uses a SAML-based method of allowing users to access online resources. Authentication is devolved to the user's organisation - the Identity Provider - which passes attributes to the Service Provider. These attributes enable the Service Provider to make authorisation decisions. Further information: <a href="http://shibboleth.internet2.edu">http://shibboleth.internet2.edu</a> The Internet2 Shibboleth group also develops software that implements the Shibboleth architecture. This software is also known as Shibboleth.</td>
</tr>
<tr>
<td>Single Sign On (SSO)</td>
<td>Provides a user the ability to input assigned authentication once and then access multiple online services.</td>
</tr>
<tr>
<td>SOAP</td>
<td>Simple Object Access Protocol. This is a definition of how to use a structured network service based on XML and HTTP to transfer data between online services.</td>
</tr>
<tr>
<td>Specification</td>
<td>A candidate for becoming a standard</td>
</tr>
<tr>
<td>SRW</td>
<td>A SOAP-based search protocol based on a subset of Z39.50 functionality.</td>
</tr>
<tr>
<td>Standard</td>
<td>A technology, format or method ratified by a respected authority</td>
</tr>
<tr>
<td>Subject Gateway / Gateway</td>
<td>A network service based on a catalogue of Internet resources. The gateways provided by RDN hubs focus on particular subject areas.</td>
</tr>
<tr>
<td>TechDis</td>
<td>TechDis provides information and advice on the use of new and existing Information and Communication (ICT), to enhance access to learning and teaching, research and administration activities for students and staff with disabilities. It is a new (February 2001) JISC service which builds on the work of DISinHE for the FE and HE sector</td>
</tr>
<tr>
<td>UDDI</td>
<td>Universal Discovery, Description and Integration. A technology for building distributed registries of Web services.</td>
</tr>
<tr>
<td>UKERNA</td>
<td>United Kingdom Educational and Research Network Association, the trading name of the JNT Association (formerly the JNT, Joint Network Team). The company contracted by JISC through the ACN to manage JANET and SuperJANET</td>
</tr>
<tr>
<td>UKOLN</td>
<td>UKOLN is a national focus of expertise in digital information management, based at the University of Bath.</td>
</tr>
<tr>
<td><strong>URI</strong></td>
<td>Uniform Resource Identifier</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td>Uniform Resource Locator.</td>
</tr>
<tr>
<td><strong>URN</strong></td>
<td>Uniform Resource Name. Use The activity of doing something with a resource that has been accessed.</td>
</tr>
<tr>
<td><strong>W3C</strong></td>
<td>World Wide Web Consortium</td>
</tr>
<tr>
<td><strong>WAYF</strong></td>
<td>Where Are You From? The Shibboleth service that provides a mechanism for routing users from a resource on their Service Provider to their point of login (Handle Service). However, it is notionally assumed to be 'optional' in the Shibboleth specification, and can be implemented either by a service provider, or as a central service, perhaps by the federation provision body.</td>
</tr>
<tr>
<td><strong>WSDL</strong></td>
<td>Web Services Definition Language. A SOAP protocol definition file for describing Web services. It enables programmers to quickly and easily support new protocols designed by third parties.</td>
</tr>
<tr>
<td><strong>XACML</strong></td>
<td>eXtensible Access Control Markup Language. An OASIS standard for the expression of access control policies. It also contains a request/response protocol, and goes some way to specifying the actual components required (such as policy decision and enforcement points) in an access control infrastructure. It is a rich, but as yet, quite obscure and underused language.</td>
</tr>
<tr>
<td><strong>XML</strong></td>
<td>Extensible Markup Language. A standards-based, electronic data format for transferring or organising information. Often used to transfer data between online services.</td>
</tr>
<tr>
<td><strong>XML namespace</strong></td>
<td>The definition of a particular use of XML used for modelling XML documents. One example is SOAP.</td>
</tr>
<tr>
<td><strong>XML Schema</strong></td>
<td>A protocol supporting distributed search and retrieve between structured network services.</td>
</tr>
</tbody>
</table>