Interoperability in Digital Libraries

Hussein Suleman

University of Cape Town, Faculty of Science

Abstract:

This chapter present the principles and practices of interoperability – the ability of systems to work together – as it pertains to digital libraries.

While there is no well-defined theoretical basis for interoperability, it has gradually emerged as a major aspect in the creation of digital library systems, particularly in modern digital repositories such as those adopted by the Open Access movement. The need for standardisation is a key element of interoperability, and is considered in tandem with the more technical elements. Principles of interoperability have emerged through experimentation and any future attempts to infuse interoperability into a system should build on these principles, such as simplicity and orthogonality. In practice, experiments with system and protocols have demonstrated what works and what does not and where there is a need for additional interventions, such as the successful OAI-PMH and RSS standards.

The key interoperability technologies currently in use in digital library systems are introduced and contextualised in terms of their applicability and motivations. In this discussion, the line between digital library standards and Web standards is intentionally fuzzy because of the increasingly symbiotic relationship between these communities.

Keywords

Interoperability, standards, protocols, components, distributed systems.

1. Introduction to Interoperability

Interoperability refers to the ability of systems to work together either to collaboratively solve a common problem or to enable the work of one or the other system. While it is frequently used in the context of computer systems, interoperability is indeed an everyday phenomenon that is taken for granted in other walks of life.

Consider, for example, typical office stationery such as a stapler or a hole punch. A stapler uses standard-sized staples – while there are usually a few choices, only a small number are readily available in any country to ensure interoperability with staplers. Hole punches are preset to make holes with spacing that corresponds to ring binders and files of a particular country. While all hole punches are uniform in one country, the standard for hole punches may be different in another country.

In an IT context, interoperability of credit cards means that a restaurant can contract with one bank to process all credit card payments made at the restaurant, irrespective of the multitude of credit cards used by its patrons. Internal electronic communication among banks ensures that the correct accounts and banks are debited and credited when a transaction occurs.

In the digital library context, the Open Access[1] movement mandates that all archives adhering to its philosophies must make the metadata for their contents accessible via the Open Archives Initiative Protocol for Metadata Harvesting[2] (OAI-PMH). This has been a primary requirement since the inception of the movement. The OAI-PMH allows for the exchange and sharing of metadata and therefore the creation of services on a level playing field based on openly accessible digital objects. In particular this has given rise to meta-archives and meta-search services.

In all of the above examples, from the generic everyday technology to the specifics of digital libraries, interoperability is an enabler that prevents monopolies, thus has a profound impact on society in general and specifically the development of online archives.

2. Concepts, definitions and principles

2.1. Definition

Wikipedia defines "interoperability"[3] as the ability of different systems to exchange data using the same file formats and protocols. This includes both those systems that interoperate for the purpose of exchanging data and those that exchange data as a consequence of communication where the exchange of data is not the primary purpose (such as X10[4] home automation controllers, which exchange data only to control devices).

---

[1]Budapest Open Access Initiative, 2002. Available online at http://www.soros.org/openaccess/.

[2]C. Lagoze, H. Van de Sompel, M. Nelson & S. Warner, *Open Archives Initiative Protocol for Metadata Harvesting*, Open Archives Initiative, 14 June 2002. Available online at http://www.openarchives.org/OAI/openarchivesprotocol.html

[3]Interoperability, Wikipedia, 2009. http://en.wikipedia.org/wiki/Interoperability

[4]Standard and Extended X10 Code Protocol. Available online at http://software.x10.com/pub/manuals/xtdcode.pdf

Lessig[5] goes further to state that "Perhaps the most important thing that the Internet has given us is a platform upon which experience is interoperable." This highlights the duality of interoperability – as both a syntactic and semantic construct. In the syntactic sense, interoperability of systems can be achieved by the exchange of data – in the semantic sense, making sense of that data in a standard manner is a more complex and often difficult task.

Syntactic interoperability is achieved using standards such as Extensible Markup Language (XML)[6], which encode data such that its structure can be understood but not its meaning. In contrast, the Dublin Core[7] metadata format is an example of a standard that focuses on semantic interoperability – standardised meaning is specified in abstract terms, with many different possible encodings.

2.2. Why Interoperability?

There are many reasons for interoperability in digital library systems. At a conceptual level, it promotes openness or choice. If an archive is able to interoperate with multiple search services, then end users may use any of the search services as a means of discovery for a single data set. This approach to search services is taken by the Networked Digital Library of Theses and Dissertations (NDLTD)[8].

Archivists typically wish to connect systems together at a service or data level so their end users may be able to search through remote collections at a single portal or access point. This single access point could be a meta-search service that is provisioned externally, exploiting interoperability to gather and harvest metadata into a meta-archive. This approach is taken by the US National STEM (Science, Technology, Engineering and Mathematics) Digital Library (NSDL)[9].

At the lower layers, interoperability results in savings in time, effort and money. Systems can be developed to use particular tools and APIs that, if standard, can be interchanged easily. Data stored and processed by such systems in standard formats will be easier to handle than proprietary formats, especially in the long term. In particular, standardisation of data formats (such as the PDF/A[10] archival subset of PDF) is usually a key facet of a preservation strategy.

2.3. Protocols, Data Formats and Standards

Interoperability frequently is defined as the standardisation of either data formats or communications protocols.

---

[5]L. Lessig, *CC in Review: Lawrence Lessig on Interoperability*, 19 October 2005, Available online at http://creativecommons.org/weblog/entry/5676

[6]T. Bray, J. Paoli, C. M. Sperberg-McQueen, E. Maler & F. Yergeau, *Extensible Markup Language (XML) 1.0 (Fifth Edition)*, W3C, 26 November 2008. Available online at http://www.w3.org/TR/xml/

[7]Dublin Core Metadata Initiative, *Dublin Core Metadata Element Set*, Version 1.1, 15 January 2008. Available online at http://www.dublincore.org/documents/dces/

[8]Networked Digital Library of Theses and Dissertations, 2009. Available online at http://www.ndltd.org/

[9]National STEM Digital Library, 2009. Available online at http://www.nsdl.org/

[10]PDF Tools AG, *PDF/A – The basics*, white paper, 22 January 2007. Available online at http://www.pdf-tools.com/public/downloads/whitepapers/whitepaper-pdfa.pdf

A standard is a specification that are maintained and endorsed by a recognised standards body – such as the HyperText Transfer Protocol (HTTP)[11], which is endorsed and maintained by the World Wide Wide Consortium (W3C)[12]. A specification is any formal statement of a data format or protocol. The advantage of using a standard is that there is some assurance of longevity and continued access as well as access by a potentially large and unconnected audience, which is crucial for many interoperability ventures.

Standards may be defined in the spirit of Raymond's Cathedral and the Bazaar[13]. They are either developed by small entities and submitted to a standards body or developed by a large community of practitioners. The Really Simple Syndication (RSS)[14] data format and the original SOAP[15] protocol fall in the former category while the OAI-PMH protocol falls in the latter category.

Data formats define the syntax and/or semantics of data used for interchange among systems. Data format standards include digital object standards - such as JPEG2000 - and metadata standards - such as IMS Learning Resource Metadata[16] to describe educational material. In addition, some data formats embed, aggregate and compose other data - such as RSS and the OAI's Object Reuse and Exchange[17].

Protocols define the communication that occurs among 2 or more parties. Typical protocols are Z39.50[18] for remote searching and OAI-PMH for metadata harvesting. Most standard protocols build on standard data formats. The OAI-PMH, for example, describes the interchange of metadata records that are themselves in standard data formats.

Thus a typical approach to building interoperable systems includes a combination of data and protocol support at syntactic and semantic levels.

2.4. Layered Interoperability

Protocols and data formats for interoperability are usually not built in isolation, but as part of a larger framework. Figure 1 illustrates how current protocols have emerged in a layered fashion, each building on a lower level of interoperability. The topmost protocols are used in modern digital library systems but have an inseparable reliance on the interoperability efforts of the Web community, which in turns relies on the interoperability of networked systems.

---

[11]R. Fielding, J. Gettys, J. Mogul, H. Frykstk, L. Masinter, P. Leach & T. Berners-Lee, *Hypertext Transfer Protocol – HTTP/1.1*, Network Working Group, June 1999. Available online at http://www.w3.org/Protocols/rfc2616/rfc2616.html

[12]World Wide Web Consortium, 2009. Available online at http://www.w3.org/

[13]E. S. Raymond, *The Cathedral and the Bazaar*, O'Reilly Media, 1999.

[14]RSS Advisory Board, *RSS 2.0 Specification*, 30 March 2009. Available online at http://www.rssboard.org/rss-specification

[15]N. Mitra & Y. Lafon, *SOAP Version 1.2 Part 0: Primer (Second Edition)*, W3C, 27 April 2007. Available online at http://www.w3.org/TR/soap12-part0/

[16]IMS Global Learning Consortium, *IMS Learning Resource Meta-Data Information Model v1.2.1 Final Specification*, 28 September 2001. Available online at http://www.imsproject.org/metadata/imsmdv1p2p1/imsmd_infov1p2p1.html

[17]C. Lagoze, H. Van de Sompel, P. Johnston, M. Nelson, R. Sanderson & S. Warner. *Open Archives Initiative Object Reuse and Exchange*, Open Archives Initiative, 17 October 2008. Available online at http://www.openarchives.org/ore/1.0/primer

[18]National Information Standards Organization, *Information Retrieval (Z39.50): Application Service Definition and Protocol Specification*, ANSI/NISO Z39.50-2003, 27 November 2002. Available online at http://www.loc.gov/z3950/agency/Z39-50-2003.pdf

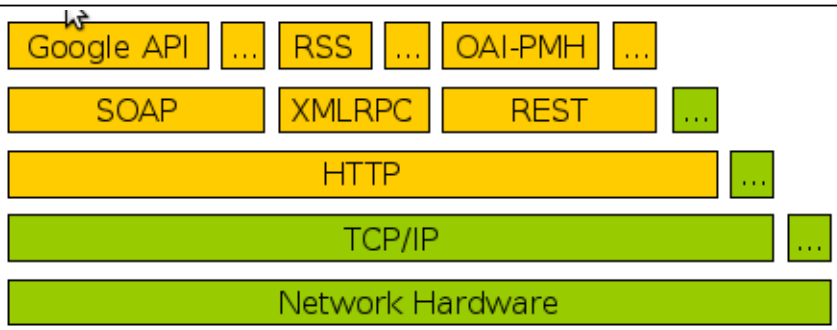

Figure 1. Layers of interoperability in a typical Web-based system

Data formats also exhibit this layering of standards. Some current digital object formats (such as JPEG2000) embed metadata in specific standard formats, encoded in XML. Some current metadata formats include support for or encapsulate other formats – such as Dublin Core[19] within RSS[20].

3. Data and metadata

Data standardisation is the most established form of promoting interoperability. Image formats such as JPEG and PNG are portable across a wide range of applications and online systems.

The Dublin Core[21] metadata standard defines 15 general elements that can be used to describe virtually anything so provides a lowest common denominator for interoperability. Other specific metadata formats, with more elements and more specific elements, exist in particular application domains. For example, the IMS Learning Resource Metadata provides a standard vocabulary and encoding to describe educational material; and VRA-Core[22] outlines how physical objects with a significant visual aspect should be described.

The rest of this chapter focuses on standardisation of protocols, as this is specific to networked systems and digital library systems in some instances.

4. HTTP, XML and Web Services

4.1.1. HTTP

HypertText Transfer Protocol (HTTP)[23] is the primary underlying protocol for data transfer on the World Wide Web, which is the common substrate for most digital library applications. HTTP defines the client-server interaction by which a client may send a request to a server and receive a document as its response. Requests are sent for

[19]Dublin Core Metadata Initiative, op. cit..
[20]RSS Advisory Board, op. cit..
[21]Dublin Core Metadata Initiative, op. cit..
[22]Visual Resources Association, VRA Core 3.0, 2009. Available online at http://www.vraweb.org/resources/datastandards/vracore3/index.html
[23]Fielding, et. al., op. cit..

documents named using a Uniform Resource Locator (URL)[24], which is a location-specific means of identifying a document based on a server name and a path within that server.

Figure 2 illustrates the format of a typical request sent by a client to a server and the format of a typical response from the server. This example only indicates the bare minimum in terms of the protocol – in a production environment, additional parameters are typically exchanged in addition to those shown.

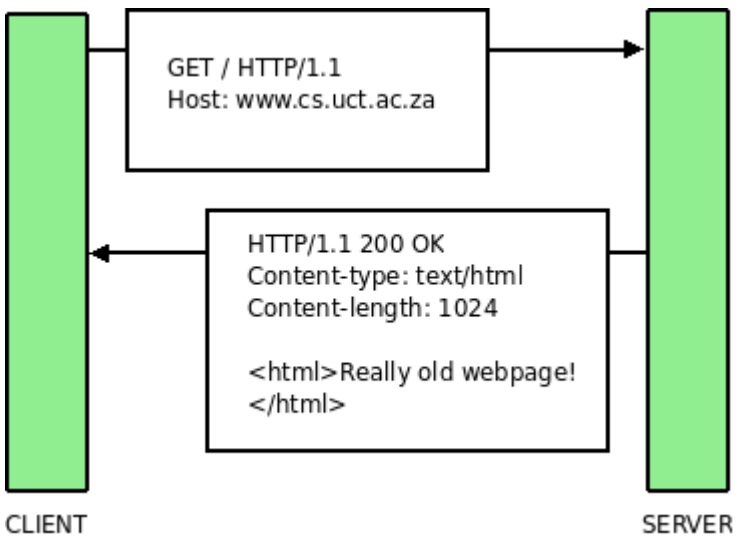

Figure 2. HTTP communication example

The request indicates the method that is being invoked – in this case GET is a request for data – as well as the URL to be used and the protocol version supported by the client. The response includes a machine-readable status code (200) and a human-readable status code (OK) in addition to the content and metadata that describes the content.

HTTP defines 7 actions or methods that may be used to communicate with a server. These are:

OPTIONS – to determine the capabilities of the server

GET – to retrieve a resource

HEAD – to retreve only the headers for a resource

POST - to submit data to the server

PUT – to insert or replace a resource

DELETE – to remove a resource

[24]T. Berners-Lee, L. Masinter & M. McCahill, Uniform Resource Locator s(URL), Network Working Group, December 1994. Available online at http://tools.ietf.org/html/rfc1738

TRACE – to trace a request as it travels through the WWW

While most HTTP clients only use a subset of these request types, the full range provides a complete mechanism for data transfer, in both directions. In addition, HTTP provides support for content type and language negotiation, date-based selective transfer, partial transfers, authentication, persistence of connections, cache and proxy control, redirection and mechanisms for reliable transfer.

When HTTP is used to handle non-static documents, the Web server will invoke an application to generate a response. These applications inspect the parameters of the request and any data attached to it; perform some processing; and assemble a suitable response. The Common Gateway Interface (CGI)[25] defines how parameters are sent to a Web-based application using HTTP. An example of a GET-based CGI request is as follows:

http://host:port/path/file?var1=value1&var2=value2&var3=value3...

The server will know how to map the URL to an application – typically mapping URL pathnames to directories on a disk. Then, the individual variables and values specified after the question mark are assumed to be the parameters for the request and are passed to the application in a manner appropriate to the programming language and environment.

POST-based CGI requests also may include large blocks of binary or textual data, so this is more suited for the case where data is uploaded to a server. Thus, using a combination of GET and POST requests, applications may be connected using HTTP as the means for transferring data.

This ability to connect together applications using a generic data transfer protocol with a wide range of features results in HTTP being a popular choice as the underlying protocol on which application-layer protocols are developed for the express purpose of interoperability.

4.1.2. XML

Extensible Markup Language (XML)[26] is a data structuring language that was derived from SGML, specifically for the exchange of machine-readable text-oriented data on the Internet. In digital library applications, XML has been used extensively to exchange structured metadata.

XML documents are primarily made up of tags and text – where the text is the actual data and the tags are the field names used to assign structure-related meaning to subsets of the text. Figure 3 is an example of a typical XML document. Each tag is surrounded by angle brackets and an optional leading slash to differentiate the start tag from the end tag. Together these tag pairs demarcate areas within the text. In this specific example, the XML represents a Dublin Core (DC) record and the tags encapsulate the individual DC fields.

[25]National Centre for Supercomputing Applications, *The Common Gateway Interface*, 1996. Available online at http://hoohoo.ncsa.uiuc.edu/cgi/

[26]Bray, et. al., op. cit..

```
<oaidc:dc xmlns="http://purl.org/dc/elements/1.1/"
xmlns:oaidc="http://www.openarchives.org/OAI/2.0/oai_dc/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
  <title>02uct1</title>
  <creator>Hussein Suleman</creator>
  <subject>Visit to UCT </subject>
  <description>the view that greets you as you emerge from the tunnel
under the freeway - WOW - and, no, the mountain isnt that close - it just
looks that way in 2-D</description>
  <publisher>Hussein Suleman</publisher>
  <date>2002-11-27</date>
  <type>image</type>
  <format>image/jpeg</format>
  <identifier>http://www.husseinsspace.com/pictures/200230uct/02uct1.jpg</identifier>
  <language>en-us</language>
  <relation>http://www.husseinsspace.com</relation>
  <rights>unrestricted</rights>
</oaidc:dc>
```

Figure 3. Sample XML document

Figure 3 is an example of a typical XML document. Each tag is surrounded by angle brackets and an optional leading slash to differentiate the start tag from the end tag. Together these tag pairs demarcate areas within the text. In this specific example, the XML represents a Dublin Core record and the tags encapsulate the individual DC fields.

XML documents also may include namespaces – prefixes for tags that allow for global uniqueness and therefore interoperable semantics.

There are many standards related to XML that provide additional facilities for system architects. In the example, the XML document includes the URL for its formal definition written in the XML Schema[27] language – this is invaluable to validate the syntactic correctness of the data using a validation engine. XML Stylesheet Transformation Language (XSLT)[28] can be used to transform XML documents from one format into another and is increasingly used for metadata tranformations. Finally, programmers may opt to use low-level APIs and tools to parse and manipulate XML – SAX[29] is a de facto specification for stream-based processing while the Document Object Model (DOM)[30] is a standard for tree-based manipulation of XML.

In typical digital library interoperability applications, XML documents, often containing metadata, are transferred from one system to another over HTTP.

4.1.3. Web Services and REST

Web Services[31] were invented to formalise this notion of connecting applications over HTTP using XML as the data interchange format – what is generally referred to as a

[27]D. C. Fallside, *XML Schema Part 0: Primer*, W3C, 2001. Available online at http://www.w3.org/TR/xmlschema-0/
[28]J. Clark, *XSL Transformations (XSLT) Version 1.0*, W3C, 1999. Available online at http://www.w3.org/TR/xslt
[29]SAX Project, *Quickstart*, 2003. Available online at http://www.saxproject.org/?selected=quickstart
[30]A. Le Hors, P. Le Hégaret, L. Wood, G. Nicol, J. Robie, M. Champion & S. Byrne, *Document Object Model Level 2 Core*, W3C, 2000. Available online at http://www.w3.org/TR/2000/REC-DOM-Level-2-Core-20001113/
[31]D. Booth, H. Haas, F. McCabe, E. Newcomer, M. Champion, C. Ferris & D. Orchard, *Web Services Architecture*, W3C, 11 February 2004. Available online at http://www.w3.org/TR/ws-arch/

Service-Oriented Architecture (SOA). The core standards in the traditional Web Services family are SOAP[32] and WSDL[33].

SOAP is a standard means of encoding parameters to an application and responses from an application in XML format. When an application wishes to communicate with another application, all data pertaining to the communication can be encoded in a SOAP message that is sent over an underlying transport such as HTTP. When the message is processed, a response can be assembled in a similar manner and returned over the same channel.

WSDL specifies the protocol for communication between 2 applications in terms of the messages that may be exchanged. A typical WSDL description includes a list of request/response pairs and the formats of the SOAP messages for each pair.

Numerous Web Services standards exist to allow for the composition, aggregation, management and discovery of services. However, given a known service endpoint (URL), WSDL and SOAP are sufficient for most current interoperability-related protocols.

REpresentational State Transfer (REST)[34] is a competing approach to SOA that defines a formal theory for the operation of HTTP. According to the REST philosophy, a request for a digital object should be encoded as a GET request in HTTP, as opposed to the SOAP approach where this maps to the POST request. The REST approach results in simpler standards that reflect the existing capabilities of HTTP in a consistent manner instead of providing a layer above it that is agnostic and unaware of the transport.

While there is no clear resolution on which SOA approach is better, both have been used widely in the definitions of different interoperability standards such as OAI-PMH[35], SRU/W[36] and RSS/Atom[37].

5. Interoperability Protocols

5.1. Metadata harvesting: OAI-PMH

The Open Archives Initiative developed the Protocol for Metadata Harvesting[38] in response to a need for a low barrier to interoperability[39]. The protocol allows for the exchange of a stream of XML-encoded records between 2 machines operating in client-server mode.

---

[32]Mitra & Lafon, op. cit..

[33]D. Booth & C. K. Liu, *Web Services Description Language (WSDL) Version 2.0 Part 0: Primer*, W3C, 26 June 2007. Available online at http://www.w3.org/TR/wsdl20-primer/

[34]R. T. Fielding & R. N. Taylor, Principled Design of the Modern Web Architecture, *ACM Transactions on Internet Technology (TOIT)*, New York, Association for Computing Machinery, 2(2), pp.115–150, 2002. Available online at doi:10.1145/514183.514185

[35]Lagoze, et. al., op. cit..

[36]Library of Congress, *Search/Retrieve via URL*, 2009. Available online at http://www.loc.gov/standards/sru/

[37]RSS Advisory Board, op. cit..

[38]Lagoze, et. al., op. cit..

[39]C. Lagoze and H. Van de Sompel, *The Open Archives Initiative: Building a low-barrier interoperability framework*, Joint Conference on Digital Libraries (JCDL), Roanoke, ACM, 17-23 June 2001. Available online at http://www.openarchives.org/documents/jcdl2001-oai.pdf

Prior to the OAI-PMH, digital archives wishing to interoperate resorted to non-standard mechanisms to transfer metadata or used federated search as a means to link together distributed systems. The latter was considered the norm but robustness and reliability were problems – with time many services tended to make changes (such as moving to a new physical machine) that impacted on interoperability. Another motivation for the development of the OAI-PMH was the realisation that those who could provide high quality services were seldom the owners of high quality data collections. Thus the OAI-PMH provides a mechanism to separate data from services and makes it possible for high quality services to easily be linked to high quality data.

The OAI-PMH defines a harvesting operation as the means to connect systems together. Harvesting refers to the transfer of collections of metadata from a source system to a target system. There is no selection or retrieval operation – the entire collection of metadata is transferred. The target system typically ingests the harvested data and indexes it in order to provide services to end users. This is in contrast to federation, where the target system formulates remote queries and submits these to the source system(s) whenever an end user makes a request for information. Harvesting is considered more robust because once the data is ingested at the target system, there is no further communication with the source system(s) – federation results in multiple points of potential failure, for each source system if there are many.

In this environment, the provider of the data is referred to as the data provider; and the system performing harvesting to provide services is referred to as the service provider.

The OAI-PMH is a client-server protocol where requests are made using URL-encoded parameters sent over HTTP. Responses are well-formed and valid XML documents conforming to a formal XML Schema. Each request is paired with a corresponding response. The underlying layer is often referred to as XMLRPC – where XML is used in the invoking of remote procedure calls. From a REST perspective, the OAI-PMH can be considered to adhere to most of the principles of REST.

The granular objects dealt with by the PMH are metadata records that correspond to an abstract notion of an item. Each item may have multiple metadata records in different formats but support for Dublin Core is a requirement. The metadata records are encapsulated within an OAI-PMH record that includes auxiliary information used to support the harvesting process. Figure 4 shows an example of an OAI-PMH record. The auxiliary information appears as a header and includes an identifier for the item, the date on which the record was last updated and a list of sets that include the record.

Sets are used to obtain a subset of the records instead of the entire set. The set name can be specified as an optional parameter when harvesting. The notion of sets is not defined globally – sets only have meaning if both the data provider and harvester have a shared understanding about the meaning of a particular set.

```
<record >
  <header >
    <identifier >oai:techreports.cs.uct.ac.za:483</identifier>
    <datestamp >2008-09-26</datestamp>
    <setSpec >796561723D32303038</setSpec>
    <setSpec >7375626A656374733D48:4835</setSpec>
    <setSpec >747970653D636F6E66706F73746572</setSpec>
  </header>
  <metadata >
    <oai_dc:dc  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
                                    http://www.openarchives.org/OAI/2.0/oai_dc.xsd"
                xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
                xmlns:dc="http://purl.org/dc/elements/1.1/">
      <dc:title >Enhancing Usability of Open Source Digital Repository Software</dc:title>
      <dc:creator >K&#246;rber, Nils</dc:creator>
      <dc:creator >Suleman, Hussein</dc:creator>
      <dc:subject >H.5 INFORMATION INTERFACES AND PRESENTATION</dc:subject>
      <dc:description >Usability of the installation ...</dc:description>
      <dc:date >2008-01-01</dc:date>
      <dc:type >Conference Poster</dc:type>
      <dc:identifier >http://pubs.cs.uct.ac.za/archive/00000483/</dc:identifier>
      <dc:format >pdf http://pubs.cs.uct.ac.za/Poster_ZA_W3_2008.pdf</dc:format>
    </oai_dc:dc>
  </metadata>
</record>
```

Figure 4. Sample PMH record

There are 6 requests, known as verbs, in the definition of the OAI-PMH. They are as follows:

Identify – get a description of the archive and its policies related to harvesting and use and reuse of metadata and data

ListSets – get a list of all the sets for which records may be requested

ListMetadataFormats – get a list of all metadata formats supported by the archive

ListIdentifiers – get a list of headers of records

GetRecord – get the record of the specified item in the specified format

ListRecords – get a list of records corresponding to the specified parameters

Figure 5 shows a typical request and response using the PMH. The request is to list all complete records in the Dublin Core format. The response includes the first batch of records.

Response

http://pubs.cs.uct.ac.za/perl/oai2?verb=ListRecords&metadataPrefix=oai_dc

Request

```
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
        http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
 <responseDate >2009-03-29T19:35:30Z</responseDate>
 <request  verb="ListRecords" metadataPrefix="oai_dc" resumptionToken="">
   http://pubs.cs.uct.ac.za/perl/oai2
 </request>
 <ListRecords >
 <record >
  <header >
   <identifier >oai:techreports.cs.uct.ac.za:506</identifier>
   <datestamp >2009-03-18</datestamp>
   <setSpec >796561723D32303039</setSpec>
  </header>
  <metadata >...</metadata>
 </record>
 <record >
  <header >
   <identifier >oai:techreports.cs.uct.ac.za:507</identifier>
   <datestamp >2009-03-18</datestamp>
   <setSpec >796561723D32303039</setSpec>
  </header>
  <metadata >...</metadata>
 </record>
 <record >
  <header >
   <identifier >oai:techreports.cs.uct.ac.za:511</identifier>
   <datestamp >2009-03-23</datestamp>
   <setSpec >796561723D32303039</setSpec>
  </header>
  <metadata >...</metadata>
 </record>
 ...
</OAI-PMH>
```

Figure 5. Sample PMH request and response – only the first 3 records are shown and the metadata records have been edited out for clarity

Due to limitations of processing systems such as XML parsers, the PMH defines a mechanism to split long lists of records into batches. When a request results in more records than can fit into a single batch as defined by the server, the records are returned along with an opaque token that may be redeemed for more records from the server – this is called the resumptionToken. The data provider and harvester continue this process of transferring partial sets of records and resumptionTokens until there are no more records to transfer.

After the first harvest has been successfully completed, a harvester may initiate further operations in the future to obtain updates. These incremental updates are based on update dates for the records. Rather than track this at the server, the client is expected to keep track of when last it harvested and thus specify a date range for future incremental harvests.

The PMH also includes features to track deleted records; deal with failures using HTTP retry mechanisms; and track the history of records in a hierarchical harvesting environment.

Currently the OAI-PMH is being widely used as a means to harvest metadata from Electronic Thesis repositories and Open Access repositories. Some other applications such as learning management systems and data curation systems also use the protocol.

5.2. Remote searching: SRU

Various protocols have been designed for remote searching in the context of digital libraries and library systems. Z39.50[40] is an established ISO standard for interoperability among library systems based on the notion of query federation. This is supported by many ILSes but not many digital repository systems, especially not those that are distibuted as open source software. While it is an accepted standard, Z39.50 has been criticised for being unnecessarily complex and for being based on outdated standards.

The Search/Retrieve via URL (SRU)[41] project has since developed a new protocol to enable remote searching. This protocol is specified in abstract terms and can be encoded into a RESTful URL or a SOAP message (previously referred to as SRW).

Figure 6 is an example of an SRU request and response. The request is specified in a query language developed for this purpose. The response includes a stream of records in any metadata format – in this example the record is the same as that used by an OAI-PMH data provider. Just like the OAI-PMH, SRU is capable of generating batches of records but each request is associated with a result set identifier so it is not stateless.

[40]National Information Standards Organization, op. cit..
[41]Library of Congress, op. cit..

Request

```
http://alcme.oclc.org/srw/search/NDL/SearchRetrieveService?
  query=dc.title%3Dopen%20AND%20dc.title%3Ddigital%20AND%20
  dc.title%3Dlibraries%20AND%20dc.date%3D2002&maximumRecords=1&
  recordSchema=default&startRecord=1&xsl=http://alcme.oclc.org/ndltd/XNDL.xsl
```

Response

```
<?xml version="1.0" ?>
<?xml-stylesheet type="text/xsl" href="http://alcme.oclc.org/ndltd/XNDL.xsl"?>
<searchRetrieveResponse xmlns="http://www.loc.gov/zing/srw/">
  <version>1.1</version>
  <numberOfRecords>2</numberOfRecords>
  <resultSetId>bn8c2q</resultSetId>
  <resultSetIdleTime>300</resultSetIdleTime>
  <records xmlns:ns1="http://www.loc.gov/zing/srw/">
    <record>
      <recordSchema>info:srw/schema/1/dc-v1.1</recordSchema>
      <recordPacking>xml</recordPacking>
      <recordData>
        <oai_dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/"
                   xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/">
          <dc:title>Open digital libraries</dc:title>
          <dc:creator>Suleman, Hussein.</dc:creator>
          <dc:description>Thesis (Ph. D.)--Virginia Polytechnic Institute and
            State University, 2002.</dc:description>
          <dc:description>Title from electronic submission form. Vita.
            Abstract. Includes bibliographical references.</dc:description>
          <dc:publisher>[Blacksburg, Va. : University Libraries,
            Virginia Polytechnic Institute and State University,</dc:publisher>
          <dc:date>2002</dc:date>
          <dc:identifier>http://scholar.lib.vt.edu/theses/available/etd-11222002-155624</dc:identifier>
          <dc:language>eng</dc:language>
        </oai_dc:dc>
      </recordData>
      <recordPosition>1</recordPosition>
    </record>
  </records>
  <nextRecordPosition>2</nextRecordPosition>
</searchRetrieveResponse>
```

Figure 6. Sample SRU request and response

5.3. Feeds and syndication: RSS/Atom

Really Simple Syndication[42] (RSS) or RDF Site Summary (RSS) is a de facto standard for the specification of lists of items related to a website. This was first used to share information about updates, especially updates to news websites, with other sites. RSS is, however, widely used for other kinds of lists, including top rated items, most accessed items, top users of a site, etc.

Unlike the previous standards that emerged from academic communities, RSS was developed to meet an urgent need and the emphasis was on simplicity. Thus, there was no XML Schema, namespace, or even formal definition at first. This resulted in many different interpretations and a slew of versions of the early specification. Most recently, RSS v2.0 has been adopted as a common baseline and there are attempts to standardise

[42]RSS Advisory Board, op. cit..

this. Atom[43] also has been defined as an alternative, more rigorous and extensible format to resolve ambiguities in RSS. The 2 formats are, however, nearly equivalent in syntax.

Both specifications define a list of items, that can be pre-generated for efficiency. This makes syndication a very desirable means of interoperability as the cost and resources needed are low. There are however tracking mechanisms such as rssCloud that allow for changes to be pushed to subscribers.

Figure 7 shows a typical RSS feed with feed-level elements and 3 items, each of which is described by a metadata record.

```
<?xml version="1.0" encoding="UTF-8"?>
<rss version="2.0" xmlns:dc="http://purl.org/dc/elements/1.1/">
<channel>
  <title>Latest posts from “Forums”</title>
  <description>Latest posts from category “Forums” on “Computer Science” board.</description>
  <link>http://moodle.cs.uct.ac.za/phpBB3/cs/viewforum.php?f=5</link>
  <lastBuildDate>Tue, 24 Mar 2009 16:58:14 +0200</lastBuildDate>
    <item>
      <dc:creator>hussein</dc:creator>
      <pubDate>Sun, 29 Mar 2009 14:32:19 +0200</pubDate>
      <guid>http://moodle.cs.uct.ac.za/phpBB3/cs/viewtopic.php?f=7&amp;t=152#p641</guid>
      <link>http://moodle.cs.uct.ac.za/phpBB3/cs/viewtopic.php?f=7&amp;t=152#p641</link>
      <title>Re: Greatest Common Divisor</title>
      <description>...</description>
    </item>
    <item>
      <dc:creator>bssste005</dc:creator>
      <pubDate>Sun, 29 Mar 2009 14:14:44 +0200</pubDate>
      <guid>http://moodle.cs.uct.ac.za/phpBB3/cs/viewtopic.php?f=7&amp;t=152#p640</guid>
      <link>http://moodle.cs.uct.ac.za/phpBB3/cs/viewtopic.php?f=7&amp;t=152#p640</link>
      <title>Greatest Common Divisor</title>
      <description>...</description>
    </item>
    <item>
      <dc:creator>bssste005</dc:creator>
      <pubDate>Sun, 29 Mar 2009 14:13:23 +0200</pubDate>
      <guid>http://moodle.cs.uct.ac.za/phpBB3/cs/viewtopic.php?f=7&amp;t=151#p639</guid>
      <link>http://moodle.cs.uct.ac.za/phpBB3/cs/viewtopic.php?f=7&amp;t=151#p639</link>
      <title>Re: Printf</title>
      <description>...</description>
    </item>
  </channel>
</rss>
```

Figure 7. Sample RSS 2.0 feed

In digital library systems, RSS often is used to indicate new items; such a feed may be integrated into external portals.

6. Validation and quality control

Quality of interoperability is measured in terms of the level of syntactic interoperability and effectiveness of semantic interoperability. The former can be computed mechanically while the latter is usually qualitative and subjective.

[43]IETF, *Atom*, RFC 4287, December 2005. Available online at http://tools.ietf.org/html/rfc4287

The OAI-PMH has multiple validation tools. The Repository Explorer[44] is interactive and helps developers during the process of writing software for systems to act as data providers. After development, the validation suite at the OAI website is used for final authoritative testing of implementations. As a result of these rigorous tests for correctness and robustness, most OAI-PMH implementations are interoperable as a consequence of adherence to the standard.

RSS and Atom feeds may be validated using the online Feed Validator[45] that checks the format of the XML.

In general, these validators check for the following:

- XML documents are well formed and adhere to formal definitions where available.
- Protocol requests can be submitted in typical sequences successfully.
- All possible errors are handled gracefully.
- Information is consistent across all requests and responses.

Validation is critical to confirm the level of interoperability of systems. Some of the more successful standards efforts have defined test suites, formal data definitions and validation tools before publicly releasing standards. Formal languages for specifying system interaction are, however, seldom used.

7. Case study: Electronic theses and dissertations and Open Access repositories

Electronic Theses and Dissertations (ETDs) are electronic versions of the traditional documents produced in paper or book format as the means of examining and documenting the contributions of a research-oriented degree. ETDs are a prime candidate for electronic archiving because there are fewer restrictions on their dissemination, unlike research articles and papers, and modern theses and dissertations are all produced electronically. As of March 2009, the Networked Digital Library of Theses and Dissertations[46], a global organisation that promotes the use of ETDs, had a collection of over 700,000 metadata records describing ETDs at various institutions around the world.

While ETDs are arguably easier to understand, deposit and manage, they are part of the bigger picture that is Open Access[47]. Open Access is a philosophy that as much information as possible should be freely accessible without artificial barriers such as subscriptions. To implement this philosophy in the context of research, all publications are either made available via publicly-accessible online journals, or copies are stored in

[44]H. Suleman, Enforcing Interoperability with the Open Archives Initiative Repository Explorer, in Proceedings of the first ACM-IEEE Joint Conference on Digital Libraries, Roanoke, Virginia, USA, pp. 63-64, June 2001. Available online at http://www.husseinsspace.com/research/publications/jcdl_2001_paper_repository_explorer.pdf

[45]S. Ruby, M. Pilgrim, J. Walton & P. Ringnalda, Feed Validator for Atom and RSS, 2007. Available online at http://feedvalidator.org/

[46]Networked Digital Library of Theses and Dissertations, 2009. Available online at http://www.ndltd.org/

[47]Budapest Open Access Initiative, op. cit..

publicly-accessible institutional repositories either before or after publication. These are often referred to as the gold and green routes to Open Access, respectively.

One common feature of all such repositories, whether for ETDs, research publications or combinations thereof, is that support for interoperability standards is a requirement. This requirement is usually that the repository acts as an OAI-PMH data provider, allowing its metadata to be harvested by remote service providers who offer services based on a meta-collection of metadata.

Figure 8 illustrates how typical ETD and Open Access archives are interconnected across universities and global service providers using the OAI-PMH. Every clear box is a source archive and every coloured box is a meta-archive. Each line represents OAI-PMH harvesting. There are 2 hypothetical countries represented with slightly different services provided in each, and a set of global services. Where possible the global services interact with the national services, but otherwise they connect directly to the source archives. The ETD collections also are presumed to be part of the Open Access collection at an international level, but are handled separately at the national level – national reporting may require tracking of ETDs produced annually while globally most researchers are interested in research irrespective of the form of the documents. Four types of institutional repository structures are depicted in the figure:

- One institution-wide repository for everything, containing OAI sets for the different types of data (Institution Y).
- One institution-wide repository for ETDs and another for Research publications (Institution W and Institution X).
- One repository in each department/unit, containing all ETDs and publications for each department (Institution A). The contents are divided into OAI sets and harvested directly by the global service providers.
- One repository in each department/unit, containing all ETDs and publications for each department (Institution Z). The contents are divided into OAI sets and harvested into institution-wide archives, before this information is shared with global service providers. The institution-wide archives act as both service providers and data providers.

While this architecture depicts complex relationships in a hierarchical but irregular system of repositories and service providers, each node needs only implement a single data provider interface or keep track of a simple list of nodes to harvest data from (or both in the case of intermediate nodes). Thus there is no single global state, which enables the creation of multiple overlapping networks of collaboration – for ETDs and Open Access in this instance.

This model, or subsets of it, has been implemented in numerous countries and contexts to loosely connect granular collections into global services. As an example, the South African National ETD project links institutions into national and global services using a system of archives similar to this.

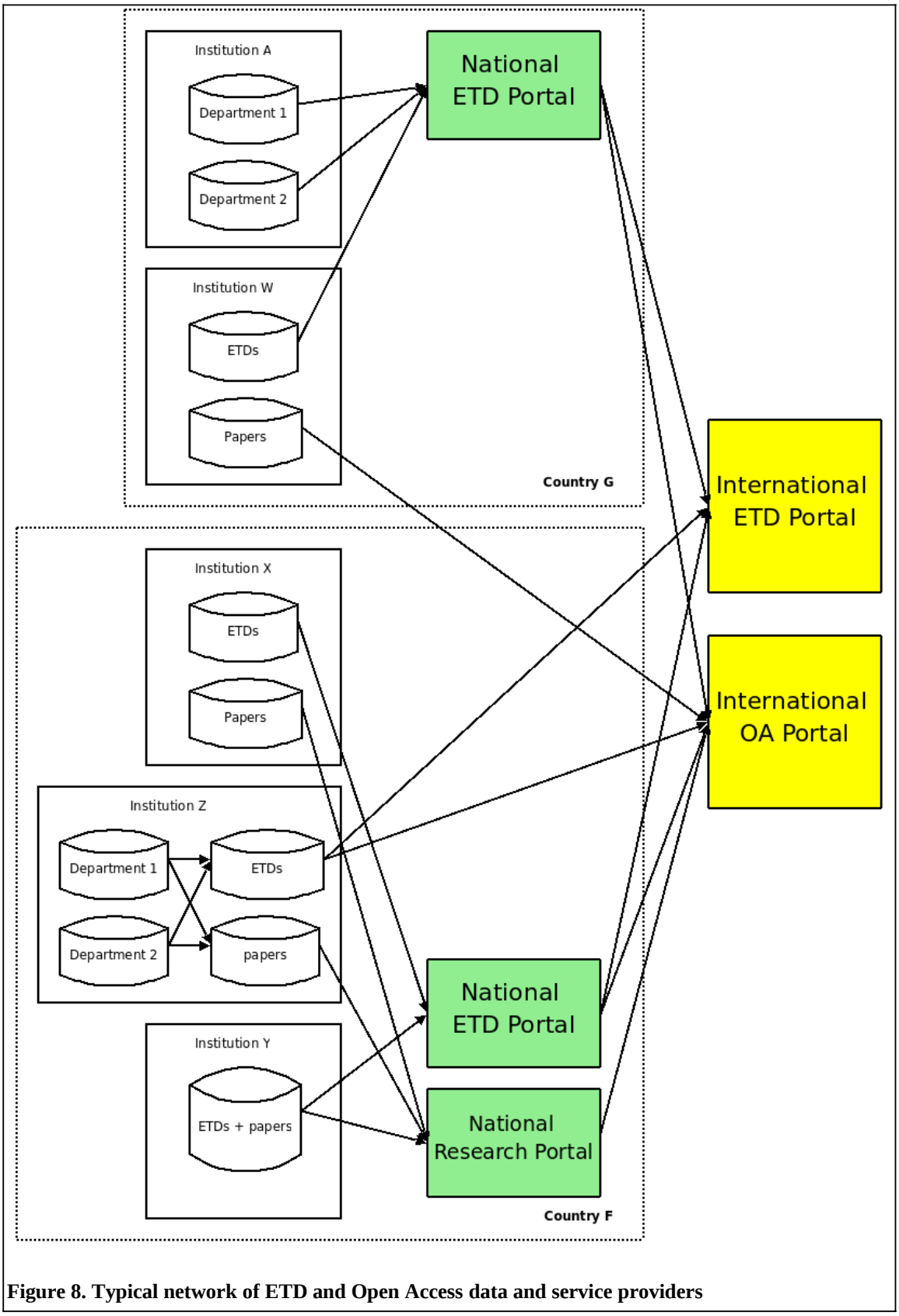

Figure 8. Typical network of ETD and Open Access data and service providers

8. Case study: Interoperability in the Developing World

In the developing world, interoperability among digital library systems is made more difficult because of the lack of resources, notably the slow or non-existent Internet connections[48]. As a result, any attempt at interoperability must take the following factors into account:

- Network connections are potentially unreliable compared to those in the North-West Hemisphere. Standards must therefore be minimal and robust, and preferably stateless and non-real-time.
- Data transfer speeds are substantially slower. Data should therefore be compressed where possible and redundancy should be factored out.
- Funding for development of systems is as scarce as network connections. Irrespective of how simple a standard is, without reference tools it will not be adopted widely.

If these concerns are addressed adequately, such as is the case in RSS and, to a slightly lesser degree, OAI-PMH, interoperability standards can be adopted globally without bias, and in fact could result in benefits for all users of the standards.

9. Emerging standards: OAI-ORE, SWORD

Emerging standards attempt to address additional needs of systems beyond metadata harvesting, remote searching and awareness. OAI-ORE[49] and SWORD[50] are such newer protocols, based on the lessons of earlier standards. Both standards define abstractions and explicit reference encodings as application profiles of existing standards (Atom and Atom Publishing Protocol respectively).

OAI Object Reuse and Exchange (ORE) is a standard that specifies inter-relationships among the constituents of aggregate and composite objects. It was developed to allow for the representation and exchange of complex objects, beyond the metadata-only harvesting enabled by the PMH. ORE does not, however, define exchange mechanisms as existing standards can be applied for this purpose.

Simple Webservice Offering Repository Deposit (SWORD) defines a standard machine interface to submit a digital object to a repository. This makes it possible for a variety of client applications to submit one or more items to any compliant repository without human intervention or with minimal intervention. Items can transferred from one repository to another or from a human user to a repository without using typical Web interfaces to repositories.

OAI-PMH + OAI-ORE + SWORD together provide a complete mechanism for one repository to transfer its contents as complex objects to another repository without human intervention, providing the possibility of ongoing synchronisation. This addresses the

[48]T. K. Thomas, India's Net connection slow, unreliable: Report , *The Hindu Business Line*, 9 March 2008. Available online at http://www.blonnet.com/2008/03/10/stories/2008031050200300.htm

[49]Lagoze, et. al., op. cit..

[50]J. Allinson, L. Carr, J. Downing, D. F. Flanders, S. Francois, R. Jones, S. Lewis, M. Morrey, G. Robson & N. Taylor, *Simple Webservice Offering Repository Deposit*, SWORD AtomPub Profile version 1.3, 22 February 2008. Available online at http://www.swordapp.org/docs/sword-profile-1.3.html

current need for object-level interoperability. Future efforts will likely explore greater interoperability at the service level.