Intermediary XML schemas

Ian Richard Mark Nansen Gartner

PhD by Prior Publication

City, University of London

Department of Library and Information Science

May 2018

Table of Contents

List of figures

Acknowledgements

I would like to acknowledge with thanks my supervisor David Bawden for guiding me through my work for this PhD by Prior Publication.

I would also like to thank colleagues past and present who made their various contributions to the research projects from which the articles presented here arose: Sheila Anderson, Jonathan Blaney, Tobias Blanke, Mark Cox, Michael Haft, Mark Hedges and Keith Jeffery.

Declaration of right to copy

Abstract

The methodology of intermediary XML schemas is introduced and its application to complex metadata environments is explored. Intermediary schemas are designed to mediate to other 'referent' schemas: instances conforming to these are not generally intended for dissemination but must usually be realized by XSLT transformations for delivery. In some cases, these schemas may also generate instances conforming to themselves.

Three subsidiary methods of this methodology are introduced. The first is application-specific schemas that act as intermediaries to established schemas which are problematic by virtue of their over-complexity or flexibility. The second employs the METS packaging standard as a template for navigating instances of a complex schema by defining an abstract map of its instances. The third employs the METS structural map to define templates or conceptual models from which instances of metadata for complex applications may be realized by XSLT transformations.

The first method is placed in the context of earlier approaches to semantic interoperability such as crosswalks, switching across, derivation and application profiles. The second is discussed in the context of such methods for mapping complex objects as OAI-ORE and the Fedora Content Model Architecture. The third is examined in relation to earlier approaches to templating within XML architectures.

The relevance of these methods to contemporary research is discussed in three areas: digital ecosystems, archival description and Linked Open Data in digital asset management and preservation. Their relevance to future research is discussed in the form of suggested enhancements to each, a possible synthesis of the second and third to overcome possible problems of interoperability presented by the first, and their potential role in future developments in digital preservation.

This methodology offers an original approach to resolving issues of interoperability and the management of complex metadata environments; it significantly extends earlier techniques and does so entirely within XML architectures.

Glossary

AIP	Archival Information Package
BRIL	Biophysical Repositories in the Lab
CASPAR	Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval
CCS	CENDARI Collection Schema
CDWA Lite	Categories for the Description of Works of Art Lite
CENDARI	Collaborative European Digital Archive Infrastructure
CERIF	Common European Research Information Format
CERIF4REF	Common European Research Information Format for the Research Excellence Framework
CMA	Content Model Architecture (Fedora)
CRIS	Current Research Information System
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DC	Dublin Core
DEFRA	Department for Environment, Food and Rural Affairs
DESIRE	Development of a European Service for Information on Research and Education
DIDL	Digital Item Declaration Language
DIP	Dissemination Information Package
DSpace	An open-access repository system developed at the Massachusetts Institute of Technology and Hewlett-Packard Labs
DTC	Demonstration Test Catchments
DTD	Document Type Definition
EAD	Encoded Archival Description
EAD3	Encoded Archival Description 3
ECHODep	Exploring Collaborations to Harness Objects with a Digital Environment for Preservation
e-GMS	e-Government Metadata Standard
EHRI	European Holocaust Research Infrastructure
ePrints	An open-access repository system developed at Southampton University
euroCRIS	European Current Research Information Systems

Fedora	Flexible Extensible Digital Object Repository Architecture, a commonly-used digital asset management architecture
Fedora CMA	Fedora Content Model Architecture
FleXML	Flexible eXtensible Markup Language
GLAM	Galleries, Libraries, Archives and Museums
ICA	International Council on Archives
ICA-ATOM	International Council on Archives - Access to Memory (software package)
ID	Identifier (XML)
IDREF	Identifier Reference (XML)
IPR	Intellectual Property Rights
ISAD(G)	General International Standard Archival Description
Jisc	Joint Information Systems Committee
LOD	Linked Open Data
MADS	Metadata Authority Description Schema
MARC	Machine Readable Cataloging
MARC Lite	Machine Readable Cataloging Lite
METS	Metadata Encoding and Transmission Standard
MICE	Measuring Impact under CERIF
MODS	Metadata Object Description Schema
NoSQL	Non SQL, a database architecture which avoids the use of relational tables
OAD	Ontology of Archival Description
OAI-ORE	Open Archives Initiative Object Reuse and Exchange
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OAIS	Open Archival Information System
OCLC	Online Computer Library Center
OWL	Web Ontology Language
PERICLES	Promoting and Enhancing Reuse of Information Throughout the Content Lifecycle Taking Account of Evolving Semantics
PLANETS	Preservation and Long-term Access Through Networked Services
PORSCHE	Performance Orientated Schema Mediation
PREMIS	Preservation Metadata: Implementation Strategies

R4R	Readiness for REF
RAE	Research Assessment Exercise
RDF	Resource Description Framework
RDF/XML	Resource Description Framework/eXtensible Markup Language
REF	Research Excellence Framework
ReLoad	Repository for Linked Open Archival Data
RELS-EXT	Relationships – External (Fedora)
RELS-INT	Relationships – Internal (Fedora)
Renardus	Academic Subject Gateway Service Europe (academic subject gateway project)
RXL	Relational to XML Transformation Language
SAN Ontology	Ontology for archival description produced by the Italian Sistema Archivistico Nazionale
Schematron	Validation language for XML which extends XML's validation criteria by incorporating assertions about patterns in instances
SGML	Standard Generalized Markup Language
SilkRoute	Publishing software for relational data in XML (AT&T Labs)
SIP	Submission Information Package
SQL	Structured Query Language
SWAP	Scholarly Works Application Profile
TEI	Text Encoding Initiative
TEI Lite	Text Encoding Initiative Lite
ToXgene	Toronto XML Server Generator
URI	Uniform Resource Identifier
W3C	World Wide Web Consortium
WGBH	Television station, user of Fedora CMA for metadata modelling
XCSL	XML Constraint Specification Language
XINTOR	Project that devised an automated method for integrating heterogeneous XML sources into a synthetic schema (2011)
XLink	XML Linking Language
XMediator	Project that devised an automated method for integrating multiple XML sources into a synthetic schema (2009)
XML	eXtensible Markup Language
XML-QL	XML Query Language

XPERANTO	Middleware for publishing relational database data as XML
XPointer	A language for referencing XML-based data or metadata components
XQuery	XML Query
XSLT	eXtensible Stylesheet Language Transformations
Xyleme	Project that devised prototype for dynamic data warehousing in XML (2001)

Chapter 1

Introduction

Intermediary schemas: the methodology and its significance

This submission for the PhD by Prior Publication details a methodology termed 'Intermediary XML schemas'. These are schemas which mediate either to others or to alternative instances of metadata conforming to themselves. They generally do not act as the final metadata container for archiving or delivery, but as mediating encoding mechanisms from which metadata in these final forms can be generated by XSLT transformations. The term 'referent schema' is used throughout this summary to describe the target schemas to which these intermediaries mediate.

The rationale behind this methodology is the resolution of certain significant problems which have been noted in employing XML architectures in digital asset management. These issues, which are discussed in detail throughout this critical summary, include the over-rigidity and lack of flexibility of the XML architecture and the impediments to discovery options which the hierarchical structures of these architectures present. Further problems arise from specific XML schemas which are over-flexible in their design and so of limited interoperability.

RDF-based models are frequently presented as alternative approaches to alleviate these problems but they themselves present significant problems for digital asset management: these include difficulties of data cleansing, IPR

issues (such as defining the boundaries of ownership) and digital preservation concerns (including incompatibility with common package-based preservation standards such as OAIS). There are, therefore, significant advantages if, as the methodology advocated here attempts to do, the most useful features of RDF, including its flexibility and the ease with which its components can be reused, may be emulated in XML.

The constituency to which this research is most relevant is that of practitioners of digital asset management, particularly those concerned with digital preservation. The methodology described here aims to enable key benefits of RDF-based metadata to be realized in XML, a robust, application-independent and well-established medium on which much digital asset management practice has been built. In particular, it should enhance digital preservation operations as it will allow these benefits to be retained within the package-based models of XML on which much preservation practice is designed.

The most significant contribution to knowledge of the methodology advocated here is the manner in which it allows some reconciliation of the conflicting requirements of flexibility and interoperability within XML architectures. This in turn enables a degree of accommodation between new developments and existing practices in metadata management by allowing rapidly changing environments to be centred on standards-based, stable cores. As a consequence, and most importantly, it should strengthen XML as a viable format for digital asset management compared to RDF-based approaches to complex metadata.

Intermediary schemas: methods

This overall methodology of intermediary schemas can be subdivided into three subsidiary methods:-

- **Method 1: t**he creation of application-specific intermediary schemas which are more constrained, easier to manage and potentially more interoperable than their referent schemas: metadata conformant to these may be converted by XSLT transformations to generate instances conforming to the referent schemas. This method is described in Article 1 (an intermediary to CERIF) and Article 5 (an intermediary to EAD).
- **Method 2:** an intermediary schema is employed as a means of serializing a template for navigating a complex XML schema by defining an abstraction of selective components from its instances. Article 3 describes the use of the structural map within the METS metadata packaging standard to serialize the semantic linkages within CERIF instances in this way; this produces a heavily constrained map of a complex application of greater interoperable potential than the raw instances themselves.
- **Method 3:** a pre-existing schema is employed as a templating mechanism or as a conceptual model in which the structure of the metadata of a compound digital object is separated from that of its constituent components and the two combined by XSLT transformations in order to generate an instance to enable its delivery. Article 2 and Article 4 describe the use of the METS structural map in

this way for the delivery of complex objects: here METS is employed as an intermediary to itself, the skeletal METS template acting as a mediator to fully-realized instances of the dissemination metadata.

The terms **Method 1**, **Method 2** and **Method 3** are used throughout this summary to refer to these three methods: the term **methodology** is used to describe the overall concept of intermediary schemas.

This critical summary proposes a synthesis of these three methods as a way of standardizing the overall methodology and enhancing its overall applicability: this is elaborated on page 78 of Chapter 6.

This submission

This submission for the PhD by Prior Publication consists of five peer-reviewed articles and this critical summary, which together define and elaborate the potential uses of the methodology of intermediary XML schemas. The articles submitted are as follows:-

- Gartner, R. (2011) Intermediary schemas for complex XML applications: an example from research information management. *Journal of Digital Information*. 12 (3) [online]. Available from: https://journals.tdl.org/jodi/Article/view/2069 (Accessed 27 February 2017). **[Article 1]**

- Gartner, R. (2012a) METS as an intermediary schema for a digital library of complex scientific multimedia. *Information Technology and Libraries*. 31 (3), 24–35. **[Article 2]**
- Gartner, R. (2012b) The digital object in context: using CERIF with METS. *Journal of Library Metadata*. 12 (1), 39–51. **[Article 3]**
- Gartner, R. (2014) Intermediary schemas and semantic linkages: an integrated architecture for complex digital archives. *International Journal of Metadata, Semantics and Ontologies*. 9 (4), 289-298. **[Article 4]**
- Gartner, R. (2015) An XML schema for enhancing the semantic interoperability of archival description. *Archival Science*. 15 (3), 295-313. **[Article 5]**

Each is referenced throughout this summary by its Article number (Article 1, Article 2 etc.). References to these articles employ the page numbers of this critical summary, not those in the works as published.

These articles were written as outputs to four research projects undertaken in the Centre for eResearch and the Department of Digital Humanities at King's College London between 2009 and 2015:-

- **Readiness for REF (R4R)** (Articles 1 and 3): a Jisc-funded project which aimed to devise new methods for facilitating data gathering in the 2014 REF exercise, particularly the use of the CERIF standard for this purpose (Centre for e-Research, 2011b)
- **Biophysical Repositories in the Lab (BRIL)** (Article 2)**:** a Jisc-funded project which aimed to create a digital library of biological

nanoimaging to enable research in live cell protein studies (Centre for e-Research, 2011a)

- **Demonstration Test Catchments (DTC)** (Article 4)**:** a DEFRA-funded project which sought to create a digital archive of environmental data in order to investigate the effects of pollution on the ecosystems of river catchments (National Demonstration Test Catchment Network, 2011)
- **Collaborative European Digital Archive Infrastructure (CENDARI)** (Article 5)**:** a European Commission-funded project which aimed to establish infrastructures for integrating European archival resources (CENDARI Project, 2013a).

I acted as Principal Investigator for the R4R project and as metadata specialist/investigator on the remaining projects.

All of the articles appeared in peer-reviewed journals. Articles 1 and 2 were included in the submission by the Department of Digital Humanities to the 2014 REF exercise (Higher Education Funding Council for England et al., 2015).

Reference is also made in this summary to one further article (Gartner et al., 2013) and one conference paper (Gartner and Hedges, 2013), both co-authored by myself, which illustrate later developments arising from the methods described within the submitted articles. These are not included in the submission as, although they do give some insight into the impact of

intermediary schemas on later research, they do not address centrally or develop further the methodology of intermediary schemas.

Literature searches for this critical summary were carried out using Google Scholar (keyword and citation searches), Library, Information Science and Technology Abstracts (keyword searches) and Web of Science (keyword and citation searches).

Research questions

This critical summary aims to address answer two overarching research questions:-

- **Research question 1:** in what ways, and to what extent, may intermediary XML schemas be employed to facilitate the management, preservation and dissemination of metadata for complex digital objects?
- **Research question 2:** How may intermediary schemas enhance metadata interoperability within XML architectures?

The discussions throughout this summary will aim to illuminate these questions and they will be addressed directly in the conclusion.

For the purposes of this critical summary, the concepts ‘digital object’ and ‘complex digital object’ are assumed to be defined in terms of function rather than materiality: following Briet’s notion of documents as enumerated by Buckland (1997, p.806; 1998, p.226), they are conceived as signs which are

organized as evidence. The latter concept may be distinguished from the former by its composition from multiple files, but it is still to be considered a unitary object by reason of its conformance to three of Briet's 'rules' for identifying a 'document' as cited by Buckland, namely intentionality, process and phenomenological position (1997, p.806; 1998, p.224): Briet's fourth 'rule', materiality, is irrelevant to defining the concept within the digital context.

Structure of this critical summary

The remainder of this critical summary takes the following form:-

- **Chapter 2** summarizes the five submitted articles and attempts to elaborate the links between them in order to establish the coherence of the research that they document. The chapter differentiates intermediary schemas by their respective functions and discusses each in turn: one category of schema enforces constraints on complex metadata architectures to facilitate interoperability and interchange (Methods 1 and 2), the other uses templating to simplify the delivery of complex digital objects and enable the flexible reuse of their components (Method 3).
- **Chapter 3** attempts to establish a context for the research documented in the articles by relating it to the research environment at the time that they were written (2011-2015) and referring to the literature of the period as required. It also assesses critically how this research carried forward or developed the work of its contemporaries.

- **Chapter 4** assesses critically the strengths and limitations of the research documented in the submitted articles. This assessment includes a discussion of the relative success of each method in the context of the project within which it originated in addition to that of the wider research environment.
- **Chapter 5** provides a critical assessment of the relevance of this research to the contemporary research environment. It examines three areas in which it is particularly relevant: the development of digital ecosystems, current developments in archival description and research relating to the application of Linked Open Data in digital asset management and preservation.
- **Chapter 6** assesses critically the potential implications of this work for future research. It examines how each of the methods may be developed further in future work, discusses the possibility of synthesizing the methods and points out how they may contribute to future work on digital preservation.
- **Chapter 7** forms a conclusion to this critical summary in which the main threads are drawn together and the research questions elaborated above are addressed and answered.
- **Articles 1-5** appear after these seven chapters and before the bibliography which lists works cited in this critical summary.

The term *schema* is used throughout to refer only to XML schemas conforming to the XML Schema Language; this is in contrast to the term *scheme* which is employed to refer generically to metadata standards and applications.

All acronyms used within this critical summary are expanded within the glossary preceding this introduction.

Chapter 2

Themes and threads in the submitted articles

Introduction

This chapter attempts to summarize the five articles submitted for this PhD by Prior Publication and to draw out the threads that link the research they document. To set them in context, reference is made to the literature cited within the articles and to other contemporaneous work.

All of these articles are concerned with XML schemas which are in some sense 'intermediary', that is they generally act as mediators to other schemas (which I term 'referent schemas'). Metadata that conforms to these schemas usually requires processing by XSLT transformations to achieve its final, often complex, delivery format. In some cases, an intermediary schema's referent may be itself: here the metadata generated by these XSLT transformations conforms to the same schema.

The articles detail the three subsidiary methods of intermediary schemas that were enumerated in the introduction to this critical summary. Methods 1 and 2 (Articles 1, 3 and 5) aim to enhance the potential interoperability (syntactic, structural and semantic) of complex XML applications by constraining over-flexible schemas in order to facilitate the management of their metadata and particularly their interoperability with other systems and applications.

Method 3 (Articles 2 and 4) also aims to facilitate the management of complex applications but does so by employing intermediary schemas as templating mechanisms; these are designed to simplify the delivery of compound digital objects and to enable the flexible reuse of their constituent components, in a manner akin to such RDF-based digital library architectural models as the Fedora Content Model Architecture (CMA).

Method 1 employs bespoke, project-specific schemas which are designed to act as mediators to their other, more established counterparts: in these articles CERIF4REF (Article 1) and CCS (Article 5) conform to this model. Methods 2 (Article 3) and 3 (Articles 2 and 4) employ the digital library packaging standard METS as an intermediary to alternatively realized instances conforming to the same schema.

This chapter introduces the three methods proposed in the submitted articles in the context of their intended functions. Methods 1 and 2 are discussed within the perspective of issues of constraint, interoperability and interchange and Method 3 within those of templating and the reusability of components.

Constraint, interoperability and interchange (Methods 1 and 2)

The use of intermediary schemas outlined in Methods 1 and 2 attempts to ameliorate frequently documented problems in the interoperability of XML-encoded data and metadata. Interoperability is variously defined as the ability to exchange and use data without manipulation (Taylor, 2004, p.369) or to use it unmodified in a variety of software applications (Schmidt, 2014). It

should ideally require no human intervention to use it in processes other than the ones for which it was created (Bauman, 2011). As Veltman points out, to achieve this effectively requires the co-ordination of semantics between systems and some degree of negotiation so that shared meaning is established before data is exchanged (Veltman, 2001, p.167).

Interoperability is more difficult to achieve than interchange, that is the exchange or transmission of metadata after some initial processing or conversion to render it comprehensible to the receiving system (Schmidt, 2014). Bauman highlights that human intervention is required to achieve interchange, whether or not this involves communication with those providing the data ('negotiated' or 'blind' interchange respectively) (Bauman, 2011). The requirement of human intervention, whether this involves developer time or data editing, inevitably inhibits the flow of data and the quantity that can be exchanged.

XML has been justifiably criticized for the limits to its interoperability, despite its credentials as a readily exchangeable format independent of any given software application. Its inherent interoperability is primarily syntactic and structural (Gradmann, 2010, p.159) (Article 4, p.138) and its standard validation techniques test only the conformance of an XML instance to the syntactic and structural requirements of a schema (Jacinto et al., 2004, p.2) (Article 1, p.94). XML and its validation mechanisms aim primarily to describe a document's structure and do not attempt to interpret the data it contains (Decker et al., 2000, p.67). Semantic control within XML is very limited: it may be achieved and validated by closed lists of attribute values or

enumerated lists of element content (Jacinto et al., 2004, p.2) (Article 1, p.94)) but is otherwise not achievable with standard XML validation techniques.

Even at the level of the syntactic and structural, interoperability in XML can be difficult to realize. McDonough casts doubt on the interoperability of structural metadata in XML because of the 'flexibility, extensibility, modularity and use of abstraction' introduced to many schemas to enhance their take-up (McDonough, 2008). Problems have particularly been found in exchanging METS documents (for example, DiLauro et al. (2005), Maslov et al. (2010)) which can be attributed to its adherence to these four principles in much of its design.

When schemas are over-flexible, extensible, modular or abstract in the ways described by McDonough, enhancing their interoperability requires some degree of constraint. He recommends, for instance, restricting the ways in which the structure of objects can be described, the use of additional schemas within a document instance, formal rules to constrain those elements of encoding an object which cannot be validated through standard mechanisms and the strict use of controlled vocabularies for otherwise uncontrolled schema semantics (such as the TYPE attribute of elements within METS) (McDonough, 2009b).

Other approaches attempt additional validation of XML instances beyond that achievable through the standard techniques of XML parsers. XCSL (Jacinto et al., 2004) and Schematron (Van der Vlist, 2007) are two commonly used

methodologies which allow the additional validation of XML instances beyond that provided by these parsers. Schematron is used, for instance, to enable the further validation of METS files in the Library of Congress's National Digital Newspaper Program (Littman, 2006). Both are useful tools but the extra validation stages required to implement them rapidly become cumbersome in the case of complex applications spanning multiple XML instances (such as CERIF, the subject of Article 1).

Method 1

Method 1 attempts to find ways to introduce additional degrees of constraint in order to facilitate greater interoperability without resorting to other validation techniques than those available as standard within XML. It does so by the design of project-specific XML schemas which represent a syntactically, structurally and often semantically constrained intermediary to a more complex and flexible established schema. CERIF4REF, designed to act an intermediary to the research information management format CERIF (Article 1) and CCS, the CENDARI Collection Schema, which can perform a similar function in relation to the archival standard EAD (Article 5), are examples of this type of intermediary schema.

CERIF4REF, introduced in Article 1, is an intermediary schema which attempts to extend the technique of architectural processing, originally a feature of SGML. This methodology allows for the construction of a bespoke, highly constrained, DTD which is mapped architecturally to an established application (such as the TEI) (Simons, 1998). The constrained DTD allows

for a small, highly focused set of elements to be constructed to meet the needs of a specific project but, when processed, for these to generate a document instance conforming to a more widely-used but overly flexible and complex standard. This technique was advocated by Simons to create bespoke DTDs for the construction of dictionaries which could then be used to generate instances conforming to the TEI (Simons, 1998) and also to convert TEI documents to object-orientated databases (Simons, 1999).

CERIF4REF attempts to emulate and enhance the functionality of architectural processing by the use of XSLT transformations. CERIF4REF is designed to mediate to CERIF, a widely-established standard for recording a variety of information required for research management (EuroCRIS, 2010). At the time Article 1 was written CERIF (version 1.3) was available as XML in 192 separate schemas that replicated the tables of the SQL database in which it was originally conceived (EuroCRIS, 2012a); a later version (1.4) simplified this architecture into a more hierarchical model better attuned to potential of XML (EuroCRIS, 2012b). CERIF4REF was designed as part of a project named Readiness for REF (R4R) (Centre for e-Research, 2011b) which was tasked with examining the compatibility of the CERIF standard with the data structures used in the Research Assessment Exercise of 2008 (Higher Education Funding Council for England, 2007); specifically it investigated its compatibility with an XML schema, RAE, produced for the 2008 exercise.

The incongruities between the two standards and the difficulties that they presented are enumerated in Article 1. Problems arose from the atomistic

approach of CERIF to data components which contrasts with the tendency of RAE to encode aggregations of these (p.96): for instance, the RAE would record the number of research assistants attached to a project, while CERIF would list them individually and aggregate them in processing. In addition, RAE and CERIF both use element content instead of internal XML ID/IDREF pairs to establish linkages between components: this requires the use of supplementary validation techniques such as Schematron or XCSL in addition to an XML parser (p.97). Designing systems to accommodate 192 interlinked schemas could potentially be difficult and complex, a problem compounded by the flexibility with which they can be applied and the extensive set of external semantics to which they must be linked if they are to function (p.93).

CERIF4REF attempts to alleviate these problems by offering a highly constrained schema specific to the requirements of the R4R project. The element set is limited to those required to generate the 19 (out of 192) CERIF instances previously identified as needed to encode the information recorded in the RAE schema (Gartner and Grace, 2010) (p.95). As is required by CERIF, metadata components are listed individually rather than by aggregation (p.95), and XML ID/IDREF pairs are used extensively to map out the internal linkages necessary to express relationships between components (p.96). Semantic components that are essential for CERIF but are not part of the metadata encoded within a CERIF4REF instance, for example the identifier of the semantic scheme used within CERIF's extensive set of linking files, are encoded in the XSLT files used to generate the CERIF instance for final delivery (p.98).

A second example of a constrained project-specific schema which mediates a problematic widely-established standard is presented in Article 5. Here the mediating schema is the CENDARI Collection Schema (CCS) which can act as an intermediary to Encoded Archival Description (EAD), the XML schema used throughout the archival community for the encoding of finding aids (Library of Congress, 2017). The interoperable potential of EAD has been questioned in the archival literature, for such reasons as its document- rather than data-centric design (Dow, 2009, p.111), the multiple approaches that it allows to encoding the same concept (Shaw, 2001, p.123) and the inconsistencies in markup that inevitably arise because of this flexibility (Prom, 2002). Proposals to alleviate these problems include regularizing tagging practices though clearer guidelines (Prom and Habing, 2002, p.172), the creation of more prescriptive schemas and the design of linked constrained standards for archival description (Shaw, 2001, p.128).

The CCS schema elaborated in Article 5 represents an attempt to devise an intermediary to EAD which is more data-like, more prescriptive and more constrained than its more established referent schema. The article, originally published in the journal *Archival Science* and so aimed at introducing the schema to the archival community, examines the facet analysis which produced the top-level components of the schema (p.156), the pervasive use of URIs to define sub-facets at fine levels of granularity (p.157), its potential for machine-readable analysis (p.160), its compatibility with the Semantic Web (p.161) and its conformance to ISAD(G) (p.162). The article also

demonstrates how CCS can function as an intermediary schema to EAD (p.166).

In its current form and using the XSLT stylesheet devised at the time the article was written, CCS can only support a constrained and relatively basic EAD instance. The CENDARI Project's guidelines for applying the schema include a mapping from CCS to EAD which demonstrates that most significant components of the latter standard are present (CENDARI Project, 2013b, p.[75]): those which are notably absent are text-centric elements such as <acquinfo>, <arrangement> or <creation>. The XSLT assumes a simple CCS instance in which basic dependencies (such as required attributes) (CENDARI Project, 2013b, p.[43]) are met but the schema's extensive set of optional components, which allow descriptions at fine levels of granularity, are not present. The article presents the possibility that the limited EAD instance that can be generated from CCS might conform to the then-planned EAD-Strict element set (Society of American Archivists, 2014), although such a limited subset of EAD has not been included in the recently published Version 3 of the standard (Library of Congress, 2017).

Both CERIF4REF and CCS demonstrate the use of intermediary schemas for constraining highly flexible and (in the case of CERIF) complex or compound XML applications in order to enable highly focused applications to be devised which nonetheless offer compatibility with widely-used established standards. They also offer the potential to enhance the interoperability of the metadata that they encode by restricting the flexibility of the schemas to which they act as intermediaries, primarily by limiting the

multiple ways in which similar concepts can be encoded: in this, they particularly implement the recommendations made by such authors as McDonough (2009b) and Shaw (2001).

Method 2

Method 2, introduced in Article 3, is a more specialized form of intermediary schema. In this case, the structural encoding features of the METS schema are used to serialize and so constrain the otherwise flexible syntactic, structural and semantic relationships of a complex CERIF application. Instead of encoding metadata within the confines of an instance of the intermediary schema itself, this approach seeks instead to encode a map of the structures of its referent metadata. This map can render the referent metadata more readily interoperable and can also form the basis of alternative realizations for final delivery which can be generated using XSLT transformations.

The article advocates this method as an alternative to using CERIF as an extension schema within METS (p.120), or as a more standard intermediary schema such as CERIF4REF (p.122), primarily in order to reconcile the divergent architectures of METS and CERIF. Here METS operates as an intermediary schema to CERIF in the sense that it acts as a template for navigating the semantic links within CERIF instances and imposes a constrained, selective map for interpreting them within an application. It performs a similar constraining function to CERIF4REF and CCS but at a more abstract level, one which encodes the structures of the semantics within

a complex CERIF application instead of its metadata content. In this way, it is something of a bridge towards the usage of METS in Article 2 and Article 4 where the schema acts as a templating intermediary for the delivery of digital content.

The mechanism by which this is achieved is to serialize CERIF linkages within the structures of elements in the METS structural map. This is done by the use of a striped syntax in which at least three nested s are used, the first of which expresses the subject of a link, the second the link itself and the third the object (p.126). This syntax is advocated by Habing and Cole as a method for implementing RDF-based OAI-ORE resource maps within the structures of a METS file (Habing and Cole, 2008). The tripartite structure in their proposal is intended to emulate the subject-predicate-object structure of an RDF 'triple', the form in which OAI-ORE resource maps are commonly represented.

The methodology advocated in Article 3 extends Habing and Cole's approach of using TYPE attributes to describe the semantics of linkages. Instead the METS <area> sub-element is used to delineate the metadata component that is referenced by its parent : this is done by the use of a simple XPointer syntax which navigates down the hierarchy of elements and sub-elements in the CERIF XML file (the diagram on Page 126 demonstrates this). The metadata itself remains within the CERIF files, the METS structural map acting instead as a navigation template for accessing this content.

This tripartite structure is designed to emulate the RDF triples commonly used to encode OAI-ORE resource maps but more complex structures than these may readily be accommodated within the METS architecture. The structural map itself may incorporate any number of nested elements to express deeper hierarchical relationships and these may be supplemented by METS <structLink> linkages which can cut across the strict hierarchies of the map. The network of relationships that can be expressed by the methodology is, therefore, potentially richer and more complex than the RDF triple on which Habing and Cole base their model.

Although the article concerns itself primarily with employing this method as a means of reconciling the divergent architectures of METS and CERIF, it has significant potential to fulfil similar functions to Method 1 in relation to interoperability and interchange. As is the case with CERIF4REF and CCS, it allows for a heavily constrained, and so potentially more interoperable, representation of a complex CERIF application to be constructed. A structural map of this type can be selective in the components that it serializes, limiting them to those required for a given application and rendering them more explicit and potentially easier to process. Using the relatively simple template of a structural map as a way of navigating the complex semantic structures within a set of CERIF files renders them more interoperable: such a template could act as a machine-actionable method of semantic navigation to enable CERIF-encoded metadata to be transferred and ingested with little, if any, human intervention.

Templating and reusability (Method 3)

The two remaining articles in this submission, Article 2 and Article 4, introduce a further application of intermediary schemas as templating mechanisms for the delivery of compound digital objects of some degree of complexity. The method proposed here employs intermediary schemas as mediators to instances conforming to the same schema, in these cases the METS standard. They act here as templates, intermediaries to fully-realized instances of themselves after their internal components are populated by XSLT transformations. The rationale behind this mechanism is to allow XML architectures to realize some of the flexibility of RDF-based approaches to representing compound objects, particularly their potential for facilitating the reuse of their constituent components.

The overarching distinction between these two approaches is detailed in Article 2 (p.101): structured packaging XML schemas such as METS or DIDL aim to create aggregations of the components of compound digital objects, usually in rigid hierarchical structures, whereas RDF models, such as the Fedora CMA (Fedora Commons, 2002), tend to disaggregate these objects into atomistic units which are recombined as required (Lagoze et al., 2005, p.130). Within a Fedora repository, for instance, compound objects are disaggregated in this way into individual datastreams which are recombined for delivery by invoking a series of RDF triples expressing their mutual relationships in the compound object. These relationships can be expressed

semantically as well as syntactically at multiple levels of granularity (Lagoze et al., 2005, p.135) (Article 2, p.101).

The advantages of the approach taken by the CMA architecture have been enumerated particularly by those who have built their repositories on the Fedora Common repository system. Lagoze et al. cite the flexible reusability of components as a key advantage: splitting a compound object into its constituent datastreams allows them to be treated as discrete units which can be recombined with ease in different contexts (Lagoze et al., 2005, p.130). This can allow, for instance, the same components to appear in multiple contexts and simple/composite objects and parent-child relationships to be neatly and elegantly expressed (Gorman and Prater, 2009). Fluid definitions of the relationships between a 'work' and its 'carrier' are also possible: the WGBH Media Library, for instance, uses the fluidity offered by CMA to express complex relationships between television news items and the video tapes on which they are physically recorded (Beer et al., 2009).

As is the case with XML schemas such as CERIF or EAD, however, this flexibility can impede interoperability. Although interoperability and reusability are cited by the designers of the CMA as one of its key rationales (Fedora Commons, 2014), they have often been difficult to achieve in practice because of this flexibility. An interoperability review of Fedora carried out at King's College London, for instance, found multiple problems associated with transferring content models across systems owing to inconsistencies in approaches to their design (Sharma, 2007).

The method described in Article 2 attempts to address some of these issues by replicating the templating functions of a Fedora Content Model within the METS structural map. The context of the article is a digital library of complex image and video objects derived from biological nanoimaging experiments (p.102). The dissemination of these objects requires the combination of images in flexible ways following models which differ according to the experiment in which their constituent images were originally captured.

Achieving this within the METS architecture as it is generally employed would be problematic. The article discusses two ways in which it can be done, one which uses the METS <mptr> element to reference a separate METS file whose constituent metadata is intended to fill a element in the structural map, the other which uses OAI-ORE aggregations to express more complex inter-component relationships (p.105). Neither is wholly satisfactory: the first is limited by <mptr>'s circumscribed place within the schema (only as a direct child of a , not, for instance, within the <par> or <seq> elements used to delineate parallel or sequential processing for components), the second by the constraints that it imposes on the use of the METS behavior and structural linkage sections (McDonough, 2009a, p.328).

The method proposed in Article 2 and extended in Article 4 attempts to emulate within the METS framework some of the flexibility of the Fedora CMA model, particularly in its potential for enabling the flexible reuse of components of compound digital objects. It does this by using METS as an intermediary template which is populated by the use of XSLT

transformations on the fly for the dissemination of a complex object. In doing this it aims, in a manner analogous to the CMA, to separate the conceptual models underlying the metadata structures of a complex object from the metadata itself (Article 4, p.141): the former is used to structure the latter (and its referent data) for delivery to the end user.

In Article 2, the delivery of a set of complex biological nanoimaging still images is achieved by encoding a model or template for each type of object within a METS structural map: a hierarchy of types and sub-types of these models can be expressed within its hierarchies (p.107). At the bottom of these hierarchies are the models for the combination of components used when a complex object is assembled and delivered; each of these contains a <fptr> (file pointer) element which would normally reference a data file listed in the METS file section. Instead of this, it here acts as a pointer to an empty <file> element, a placeholder for digital content which is only realized when the METS template is processed by XSLT for delivery. Small subsidiary METS files, essentially containing only lists of the component files that need to be assembled into the compound object for delivery, are used to populate these empty <file> elements during XSLT processing.

An even more skeletal METS template is used to generate video files where the exact number of images required will vary from sample to sample and so cannot be modelled directly in the structural map. In this case both the structural map and the file section act as placeholders only and the templating model is recorded within the subsidiary file for each video (p.109). In both methods, the METS behavior section is used to record the XSLT stylesheets

necessary to generate the METS file in its fully-realized form and to identify any software needed to process and render the final object for delivery.

A similar method of employing METS as a templating mechanism is proposed in Article 4 although here a higher-level structure for defining conceptual models and their relations is employed which allows for some degree of recursion. The article advocates an integrated XML architecture for a complex archive of environmental data in which the use of METS as an intermediary templating schema is complemented by the use of METS and MADS schemas for the construction of semantic concept maps. The templating function of intermediary schemas is discussed in section 4 (p. 140) which examines their use in defining conceptual models for data types and user *personae*.

This article introduces a new layer of abstraction to the definition of templates or 'conceptual models' as they are here called (p.140). The conceptual model for each data type is defined, as in Article 2, within a METS file's structural map. This is then invoked for delivery by a higher-level METS file which uses an <mptr> element to associate it with a corresponding to its data type. As in Article 2, the METS behavior section links each desired realization of these intermediary schemas with the tools (specifically the XSLT transformations) required to produce it. The same device is also used to associate multiple *personae* to a data type in order to allow varying input mechanisms for different classes of users within the heterogeneous community of data providers for this archive. A skeletal representation of this architecture is shown in the article's Figure 1 (p.142).

Using the <mptr>element in this way allows for a greater degree of abstraction than the method described in Article 2 alone. The METS file here records the highest-level architecture of an application, the second-level intermediary schemas of Article 2 the detailed content models for compound objects, and third-level files the constituent components of these objects. This tripartite model enables a separation of structure, metadata and data, all of which can be combined and reused as necessary in a manner analogous to the flexibility of such models as the Fedora CMA.

Conclusion

The five articles summarized here demonstrate the applicability of intermediary schemas to two primary functions, the first for constraining complex and flexible XML applications in order to enhance their interoperability (syntactic, structural and semantic), the second as a templating mechanism to allow conceptual models to be created for complex applications which separate their metadata structures from the metadata of their constituent objects. Both functions attempt to reconcile the potentially contradictory requirements of interoperability and flexibility within XML architectures.

Constrained, project-specific schemas such as CERIF4REF and CCS, which revive the ideas behind SGML architectural processing, have the potential to constrain and so render more interoperable problematic schemas, such as CERIF and EAD respectively, while retaining the advantages of

compatibility with these community-established standards. A similar degree of constraint, which avoids the use of bespoke schemas, may be achieved by the technique of using the structural map of the METS schema as a template for the navigation of semantic links within a referent schema by explicitly serializing their relationships.

The second function of an intermediary schema, as a templating mechanism for encoding a conceptual map of a complex application, offers a viable alternative to RDF-based architectures, such as Fedora CMA, which disaggregate the components of complex objects for flexible recombination on delivery. Using these schemas as mediators to fully-realized instances conforming to themselves allows the metadata structures for a complex object to be separated from the metadata itself and for its constituent components to be recombined and reused with a high degree of flexibility. Complex structures can readily be encoded using mechanisms which allow multiple METS files to act as intermediaries, each at a differing level of abstraction.

Although conceived for a variety of projects, the intermediary schemas documented here all address similar issues, how complex and (from a systems design standpoint) messy metadata can be rendered tidier, and so easier to design applications for, without forgoing the expressive power required of it. As these articles often stress, the approaches advocated here are only initial attempts at addressing these issues: further research (some ideas for which are discussed later in this critical summary) will be needed further to realize their potential.

Chapter 3

The research in context

Introduction

This chapter attempts to place the research discussed in the submitted articles within the overall context of preceding or concurrent work and also to assess critically the extent to which it extends or enhances this work. It aims to do so by reference to the literature that precedes or is contemporaneous with the articles.

The first section examines the intermediary schemas of Method 1 in the context of previously established techniques for employing mediating mechanisms between metadata schemes. The second relates the serialized metadata structures of Method 2 to previous work that attempted similar approaches to handling complex, disaggregated metadata. The final section relates the use of the METS schema for templating or conceptual modelling of Method 3 to previous work and attempts to assess the extent to which it represents an original advance on its antecedents.

Method 1

By function and, to a certain extent, technique, Method 1 can legitimately be compared to previously established approaches to enabling metadata interoperability and standardization. Chan and Zeng's much-cited study (159 citations on Google Scholar) (2006) highlights a number of techniques for interoperability at the schema level with which intermediary schemas may appear to be congruent; those most relevant are:-

- ‘crosswalks’, ‘semantic and/or technical mapping[s] (sometimes both) of one metadata framework to another metadata framework’ (Patel et al., 2005, p.21), one-to-one mappings such as EAD to MODS (Bountouri and Gergatsoulis, 2009) or Dublin Core to MODS (Library of Congress, 2012)
- ‘switching-across’, synthetic switching schemes, often based on unqualified Dublin Core, to which multiple metadata schemes can be mapped (Chan, 2005); these include the OAI-PMH schema for metadata harvesting (Lagoze and Van de Sompel, 2015) and schemes that formed the core of OCLC’s Metadata Switch (Godby et al., 2003) and Crosswalk Web (Godby et al., 2008) services
- ‘derivation’, a new schema derived from an existing one, usually in a simplified form, which retains a similar basic structure to its source schema but allows for variations in depth and detail (Chan and Zeng, 2006): examples of these include TEI Lite (Burnard and Sperberg-McQueen, 1995; Burnard and Sperberg-McQueen, 2006), MARC Lite (Library of Congress, 2008) and CDWA Lite (Stein and Coburn, 2008)
- ‘application profiles’, schemes which synthesize terms from diverse vocabularies or other schemes and are customized for specific applications (Heery and Patel, 2000): when initially conceived, although not always applied later, application profiles were designed to enhance interoperability by imposing restrictions on a base standard, so allowing any system that can handle this base to understand a profile derived from it (Duval et al., 2006, p.242). Examples of these include the early DESIRE project (UKOLN, 2000), the DC Library Application Profile (Dublin Core Metadata Initiative, 2004), the e-Government profile e-GMS (UK Government, 2006), the Scholarly Works Application Profile (SWAP) (Allinson et al., 2007) and the Renardus project profile for subject gateways (Neuroth and Koch, 2001).

In a very basic sense, intermediary schemas may be considered forms of crosswalks or switching schemes, as they provide a mechanism for mapping between schemas, that is 'the process of associating elements of one set with elements of another set' (Semantic World, n.d.). They could also, and more accurately, be described as a variant of derived schemas as they are designed specifically to constrain a single established schema, usually of considerable complexity and flexibility. They might also be seen as at least partially congruent with the technique of application profiles, which similarly construct their element sets from outside schemes, although, unlike derived schemas, usually from more than a single source.

Although these four techniques for schema interoperability offer a useful framework for initial considerations of the context within which intermediary schemas can be located, there are significant ways in which the methods proposed here differ from or extend these established approaches. The following could be highlighted as the most important of these:-

- The relationships of intermediary schemas to their referent schemas are more fluid and complex than those of derived schemas, crosswalks or switching schemes: in many cases their elements do not have a one-to-one match with their counterparts in their referents nor do they attempt in most cases to replicate their basic structures (as derived schemas do)
- Intermediary schemas do not follow the 'mixing and matching' (Heery and Patel, 2000) approach of many application profiles: they do not attempt to create new schemes which extend beyond the semantic boundaries of their referent schemas, as application profiles often do (for instance the DC Library Application Profile (Dublin Core Metadata Initiative, 2004) which

incorporates additional elements to DC). They are also subsidiary to their referents, unlike application profiles which are often intended to act as independent schemes

- Intermediary schemas offer more sophisticated mechanisms to accommodate incongruities between schemas where no one-to-one semantic match exists between elements (for instance, the CERIF element <cfOrdUnitId> in Figure 8 of Article 1 (p.98), which has no equivalent in RAE). Associating an intermediary schema with an XSLT transformation as an integral complement allows the relatively simple construction of synthetic elements which can incorporate metadata components already present in but distributed throughout the source schema and additional ones created during processing
- Intermediary schemas provide a mechanism for reconciling divergent schema architectures, for example, the aggregated approach of the RAE schema and the disaggregated methods that predominate in CERIF (Article 1, p.95): the XSLT transformation provides a mechanism for this reconciliation by allowing aggregations, counts and other syntheses to be incorporated into the realization of the intermediary schema to its referent. This mechanism is, however, limited to translation from more specific, disaggregated, metadata to its aggregated form
- Intermediary schemas allow descriptions at more specific levels of granularity than may be allowed in an established schema, for instance in the more data-centric architecture of CCS as compared to the document-centric focus of its referent schema EAD (Article 5, p.152): once again, the XSLT transformation acts as a mediator between divergent granularities, although, as is the case in the bullet point above, this mediation works more readily in

translating from the more to the less specific, from the data-like components of CCS to the more flexible document-like structures of EAD

- The constraining mechanisms of intermediary schemas also extend beyond those of derived schemas and application profiles (at least when these follow their early guiding principle of restriction). When attempting to impose constraints, these two methods are limited to reducing element sets to a more confined semantic environment than their sources. CERIF4REF, by contrast, uses such techniques as enumerated lists or ID/IDREF pairings to enable constraints on an application which extend well beyond those of a employing a more confined element set alone.

From the above discussion, it seems clear that intermediary schemas find only a relatively limited place within these traditional models and certainly extend them in a number of distinctive ways. In addition to these approaches, a very small number of projects that preceded those documented in the submitted articles have adopted approaches in which they created bespoke XML schemas explicitly labelled 'intermediate', 'intermediary' or 'mediating'. The following discussion highlights those documented in the literature up to 2011, the date of Article 1, which most closely resemble the intermediary schemas in the submitted articles.

Some attempted schema mediation by creating synthetic schemas into which sources could be merged: these include an XML data integration system proposed by Almarimi and Pokorny which utilized a common 'global' schema created by the merger of source schemas (2005, p.26), the PORSCHE project which constructed an 'intermediate mediated schema' by merging input schemas (Saleem et al., 2008, p.638) and the XMediator and XINTOR projects which aimed to automate the integration of diverse XML sources into mediated schemas (Nguyen et al., 2009; Nguyen et al., 2011).

Others used intermediate XML schemas for data integration, for instance, the National Institute of Health which employed 'a common XML intermediate' (Shaker et al., 2002, p.693) for merging XML query results from heterogeneous sources or the Xyleme project which used such schemas to host abstract models of heterogeneous XML data sources in order to enable the construction of views across them (Abiteboul et al., 2001, p.46; Abiteboul et al., 2002, p.238; Cluet et al., 2001, p.273). A further application used them as nodes within a peer-to-peer data management system mediating between XML and OWL sources (Halevy et al., 2003, p.764). Another employed them as a conversion mechanism from XML to ontology instances (Bohring et al., 2005, pp.150–151); here the schemas were generated from source instances to enable the definition of an OWL ontology to which instances could be mapped.

All of these projects are essentially mapping mechanisms, most of which follow the model of switching across (particularly Abiteboul et al. (2002), Halevy et al. (2003), Almarimi and Pokorny (2005), Saleem et al. (2008), Nguyen et al. (2009) and Nguyen et al. (2011)). They do not attempt the constraining functions of CERIF4REF or CCS, nor, as they function as mediators between a number of heterogeneous schemes, do they share their integral relationship with a single scheme (even if this scheme is embodied in multiple schemas). They do not attempt any reconciliation of disjoint architectures as CERIF4REF endeavours to do in translating between the aggregated components of RAE and their disaggregated equivalents in CERIF. Nor do schemas derived from source instances, such as in Bohring et al. (2005), achieve the level of abstraction of a Method 1-type intermediary schema which allows them to enable this reconciliation.

To sum up, while intermediary schemas such as CERIF4REF and CCS can be aligned in a basic sense with some established approaches to schema interoperability, such as derived schemas or application profiles, they differ in significant ways in terms of the complexity of types of mediation that they can facilitate and the techniques that they

employ to enable this. They also differ significantly from other approaches to employing 'intermediate', 'intermediary' or 'mediating' XML schemas that preceded them in their overall purposes (constraining their referent schemas and reconciling divergent architectures) and the mechanisms that they deploy.

Method 2

Method 2, in which a complex set of semantic relationships in a diverse set of XML metadata instances is structured by serializing it within the METS structural map, has no direct analogue in the preceding or contemporaneous literature. Similarly, the primary technique employed for encoding these serializations, using the XPointer syntax within the BEGIN attribute of the METS <area> element to address the location of metadata components, is also without precedent.

It is valid, however, to compare this method with mechanisms for modelling the internal structures of compound objects in such packaging standards as OAI-ORE (Open Archives Initiative, 2008; Open Archives Initiative, 2011) or METS as it is conventionally employed. Both the OAI-ORE resource map and the METS structural map, when used to describe logical structures such as the internal contents of a compound object (a digitized book, for example), attempt to impose a constrained representation of its referent object and to define a single view of it in preference to others that could be conceived. It then maps pathways to its constituent data objects so allowing the construction of a synthetic version of the representation for delivery.

The rationale for this method also bears comparison with the Fedora data model, particularly its 'Relations datastream' (RELS-EXT and RELS-INT) (Fedora Commons,

2002; Lagoze et al., 2005, p.135), the set of RDF statements for expressing inter- and intra-object relationships within a repository. As is the case with OAI-ORE and METS, though with more flexibility, a constrained view of the universe of possible presentations of the datastreams underlying compound objects are encoded in these structures in order to facilitate their delivery.

All of these approaches to defining 'maps', which impose a logical structure on their respective compound objects and constrain the ways in which their constituent components can be presented, deal with data objects; the approach taken in Article 3 attempts to treat *metadata* objects in an analogous way, as components which need structuring and constraining into logical views for dissemination. It is this attempt to treat metadata structures as a complex compound object, and map them in a manner analogous to these preceding approaches to mapping data structures, that this method is original.

The technique of striped syntax employed to do this has only one recorded prior use in the literature, that of Habing and Cole's (2008) suggested method for serializing OAI-ORE resource maps in the METS structural map. Habing and Cole's use of the syntax is slightly different to that advocated here: they employ the LABEL attribute of elements to express the semantics of each component in a 'triple' whereas Method 2 extracts these semantics from the CERIF XML instances by their respective XPointer references.

Method 3

The use of the METS structural map for templating or conceptual modelling (Method 3) has very limited precedents in previous research or practice.

One category of earlier work which could be considered an antecedent is a small number of projects that have attempted to establish methods for defining templates within XML architectures in order to support middleware between relational databases and XML. A notable early example was XPERANTO, a piece of middleware for creating XML documents from SQL databases using queries written in XML-QL (Deutsch et al., 1998) (a now obsolete predecessor to XQUERY) (Carey et al., 2000a; Carey et al., 2000b; Fan et al., 2002). A similar middleware project from the same period, SilkRoute, also used a templating approach to interface between relational databases and XML, in this case using RXL, a declarative query language designed to query databases and add required tagging in order to generate a valid XML document (Fernandez et al., 2000; Fernandez et al., 2001, p.12).

Other approaches that could be considered precedents attempted to redefine the XML Schema language to incorporate templating features. A small number of projects made use of FleXML, a short-lived extension to XML which relaxed some of its restrictions in order to allow 'cocktail schemas' to be created in which placeholders (processing instructions) could fill in fragmentary XML documents on delivery (Parekh et al., 2006, p.153). This was used in a joint project by Columbia University and the United States Air Force Research Laboratory to design autonomic (self-managing) software (Kaiser, 2004, p.9). A further attempt to extend the XML syntax to introduce templating features was ToXgene, a tool designed to create large corpora of synthetic XML documents for use in benchmarking and similar tests; this again extended XML Schema to incorporate

features of templating components such as constraints on string length or minimum or maximum numbers of occurrences (Barbosa et al., 2001, 2002, p.618).

The METS-based methods proposed in Articles 2 and 4 differ in significant ways from these antecedents. The principle advance that they represent is that they allow templating to occur within the XML environment using only standard techniques such as XSLT transformations. Unlike FleXML and ToXgene, they do not require any changes to the XML Schema language itself, which would severely curtail the interoperability of the METS templates that they employ. Nor are they contingent on any middleware, as is necessary for XPERANTO and SilkRoute documents to realize their associated instances.

Method 3 also differs fundamentally from this earlier work in that it generates the metadata from which the final digital object is assembled for delivery, not the object itself. The methods described above generate final document instances, usually employing relatively skeletal XML structures which solely translate their referent data into a new medium for dissemination. XPERANTO and SilkRoute, for instance, are publishing mechanisms for relational data and the Columbia University project is intended primarily to create synthetic documents for benchmarking.

Because the techniques of Method 3 operate as templates to metadata, they work at one level of abstraction from the instances that are eventually generated: this allows them to define conceptual models rather than simple templates alone. This abstraction enables them to offer the potential to introduce much greater degrees of complexity into these models, by, for instance, incorporating the levels of recursion described in Article 4 or by utilising the hierarchies of the METS structural map and the facility of METS structural links to cut across these. Features such as these enable them to emulate the

flexibility of architectures such as Fedora CMA in a manner in which the antecedents described above cannot in anything more than a rudimentary fashion.

There is no mention in the preceding literature of approaches to employing the METS structural map for templating and the definition of conceptual models in a similar manner to that proposed in Articles 2 and 4. The only approach previously taken to enable aggregations and component combinations within the structural map is the use of the <mptr> element described in Article 2 (p.105): this acts a placeholder that references a separate METS file containing metadata on the content represented by its parent element.

Three preceding projects in the literature made some use of <mptr>. The ECHODep project, which examined issues of digital repository and preservation interoperability, used it to record iterations of digital preservation packages as they are ingested and disseminated in repositories (Ingram, 2009; Unsworth and Sandore, 2010). A project at New York University, which attempted to use METS for the archival preservation of websites, employed <mptr> to reference alternative captures of a site (Myrick, 2004, p.38) and as a method of mediating between varying layers of abstraction from a website as a conceptual entity, to daily captures of its content, to single HTML pages (Guenther and Myrick, 2006, p.159). The third was an early digital archive designed by OCLC in which object-level METS documents in a DIP were referenced from a higher-level manifest METS file by the use of this element (Goodkin, 2004, p.16).

The models of Method 3 may be built using the full architecture of the METS structural map and structural links sections and so allow for significantly more sophisticated and complex templating and modelling than is possible by employing the <mptr> element alone. It also allows more flexibility of placement within the structural map, allowing components to be referenced from anywhere that an <area> element can be used, for

instance within <par> and <seq> elements to define parallel or sequential processing (Article 2, p.107); this is more flexible than the <mptr> element which is limited to placement as a child of a .

Method 3 also separates data and metadata more rigorously than using <mptr> which, because it references an entire METS document, usually entails their conjoint packaging. The subsidiary METS files in Method 3 are file lists with minimum additional metadata (usually only a skeletal structural map to ensure their validation). This separation allows greater flexibility in the design of the templates or conceptual models encoded in the primary METS file and the easier reuse of the digital objects referenced in the subsidiary files. In this way, this method moves some way beyond the use of <mptr> for aggregation.

The application of Method 3 in Article 4 does employ the <mptr> element to allow linkages to be made between a high-level abstract conceptual map (in this case of data types) and lower-level templating METS files of the type described in Article 2. This approach does bear similarities to that of Guenther and Myrick (2006) but here <mptr> is only employed at the higher levels of a hierarchy of abstraction, not throughout the architecture as a whole. At these more abstract levels, <mptr> can usefully populate an entire but its limitations impede the flexibility and granularity required of templates or conceptual models at the lower levels of such a hierarchy.

Conclusion

This review of the context within which the techniques discussed in the submitted articles can be considered reveals that, although there are some antecedents for these

approaches, they extend earlier work in significant ways. The intermediary schemas of Method 1 have some precedent in techniques such as crosswalks, switching across, derivation and application profiles but take them forward in terms of the fluidity of the relationships between intermediary and referent, in particular their ability to reconcile architectural and semantic incongruities and their usage as constraining mechanisms.

Similarly, the serialization techniques of Method 2 bear some similarity to previous uses of METS, OAI-ORE or the Fedora CMA for representing the internal structures of compound objects, but extend these to a new level of abstraction by modelling the structures of metadata instead of those of the object data itself. The striped syntax employed has been deployed previously but is here extended to extract semantics directly from the referent metadata instead of recording them directly.

Using the METS structural map for templating or the definition of conceptual models (Method 3) has some remote precedents in earlier work but, unlike these earlier projects, does not require dedicated middleware or changes to the XML schema language. Unlike this earlier work, the techniques advocated in these articles generate metadata not the digital object itself and so introduce a higher level of abstraction which can facilitate more complex modelling.

Chapter 4

The strengths and limitations of these methods

This chapter attempts to assess critically the strengths and limitations of the methods advocated in the submitted articles, each of which is discussed in turn. Where appropriate, the success (or otherwise) of each method within their originating projects is discussed in addition to a discussion of their realized or potential strengths and limitations in a wider context.

Method 1

The prime rationale for employing intermediary schemas as constraining mediators to referent schemas (Method 1) is to facilitate greater interoperability between complex, flexible and often disparate sources and so to enhance the possibilities for the interchange and transmission of their metadata. One of the primary strengths of the use of schemas of this type should be their capacity to facilitate this interoperability in practical environments in which such problems of complexity and heterogeneity are present and these problems originate from over-flexible schemas. The referent schemas discussed in the articles, CERIF (for research information management) and EAD (for archival finding aids), are two which are established standards in their respective domains: others could readily be added to their number, particularly in such areas as text encoding in the humanities where the TEI presents similar issues and problems.

There is evidence from the Readiness for REF (R4R) project that CERIF4REF demonstrates this capacity, as it proved a viable mediating format for several diverse sources of research information management metadata. It proved possible to populate

CERIF4REF with metadata from the King's College London CRIS Research Gateway (Cox, 2011, p.16) and to establish its compatibility with a number of test repositories (Cox, 2011, p.17), although an intended aim of exchanging CERIF4REF-encoded metadata between institutions could not be achieved owing to confidentiality issues (Russell, 2011, p.3).

In addition, CERIF4REF formed the basis for a number of plugins for well-known repository systems developed as part of the project. These plugins, constructed for ePrints, DSpace and Fedora, all successfully generated CERIF4REF instances for interchange from their respective repositories at Southampton University, Edinburgh University and King's College London (Russell, 2011, p.3).

CERIF4REF also proved itself sufficiently extensible to accommodate changes in the CERIF standard that post-dated its initial development. One of the most significant of these was an extension developed by the MICE (Managing Impact under CERIF) project to incorporate a new model for describing research impact (MICE Project, 2011). This project introduced extensions to CERIF to incorporate impact statements and their accompanying indicators and measures into its architecture (Gartner et al., 2013, p.469). The model presented in these extensions was of considerable complexity but CERIF4REF was readily extended to accommodate it (Gartner et al., 2013, p.474).

There is some evidence that this intermediary schema method has gained some acceptance outside these projects as a viable approach to enhancing interoperability within complex and disparate metadata environments. Two articles in the literature which cite Article 1 recommend its adoption. One suggests employing a similar method for data sharing within a paediatric oncology research network: a proof-of-concept study for such a network employed, in a similar way to CERIF4REF, a core XML data model and XSLT transformations for mediating between a number of disparate data

sources (Hochedlinger et al., 2015). A further article advocates defining 'common intermediary XML schemas for complex applications, in interoperable semantic and syntax contexts' as a means of enabling metadata interoperability at flexible levels of granularity within an open-access research product repository (De Biagi et al., 2012, p.88).

The CCS schema introduced in Article 5 has yet to be implemented practically in a working environment and so cannot claim the same empirical evidence of practical usability as CERIF4REF. It was not implemented as the basis of the CENDARI project's archival directory as this employed the International Council on Archives' ICA-ATOM software which could not be adapted to incorporate this new schema. Because of this, claims to the strengths of CCS as a mediator to EAD must remain at the level of assessing its theoretical efficacy but they may nonetheless viably be made.

The principle strength of CCS is its more data-centric approach than that of EAD; this has the potential to alleviate some of the interoperability problems associated with employing the more established standard (as noted, for instance, by Shaw (2001)). These have so far hampered the creation of union catalogues of archives which can equal the degree of metadata depth and sophistication that MARC-based catalogues can offer. Owing to the modelling of EAD on the form and structure of document-based finding aids, archival union catalogues, such as the UK National Archives' *Discovery* service (The National Archives, 2015), have generally been limited to offering free-text searches or those limited to a small set of fields. The more semantically constrained and data-like structures of CCS should offer opportunities to enhance their functionality significantly.

The enhanced possibilities for interoperability presented by CCS, and the finer levels of granularity at which this may be achieved, also offer the potential for new approaches to

interacting with the archival record in a more insightful way than has previously been possible. The dynamically-constructed research guides posited in Article 5 (p.160) are one example of how CCS could offer new approaches to treating the archival record as a more malleable resource capable of delivery in ways that move beyond the traditional finding aid.

The strengths of intermediary schemas as constraining mediators to established referent schemas should also be considered in the context of their limitations. One significant counter-argument that could legitimately be made to their use is noted in Article 3 (p. 124), namely that a profusion of bespoke, unstandardized schemas is likely to nullify the benefits that they offer for enhancing the exchange of information. This could become problematic if, as is the case for the plugins devised for the R4R project, systems are increasingly built on the intermediary schemas and not their more established referents. The possibility could then arise of reduced interoperability as isolated systems, each employing a bespoke intermediary schema, fail to communicate with each other.

Intermediary schemas of this type also have two important limitations in comparison to architectural processing, the earlier method that they seek to emulate. The first, pointed out in an early article on architectural processing and XML, is that they do not allow validation against both schemas as part of processing (Lubell, 2001, p.409): each transformed instance needs to be validated on creation, a potentially time- and process-consuming function. Secondly, the separate components required for intermediary schemas, the XML schema itself and its corresponding XSLT stylesheet, can become unsynchronized unless managed very carefully, a problem which cannot arise in architectural processing where the templates for the transformations are integrated into a DTD or schema (Lubell, 2001, p.409).

These limitations need not invalidate the strengths of the intermediary schema approach enumerated above. Although the validation of instances against both schemas as part of architectural processing is undoubtedly more unwieldy than the validation of each individually, the speed of current validation tools is unlikely to render this problematic in practice. Issues of the synchronization of an XML schema and its associated XSLT stylesheet can be resolved by well-established techniques of version control. The problems of an over-proliferation of bespoke intermediary schemas may also possibly be alleviated by the adoption of techniques from Method 2 which are discussed in the next section. These limitations must be considered in the design of systems that employ intermediary schemas although they do not present insuperable barriers to their adoption.

Method 2

The use of the METS structural map to serialize metadata structures (Method 2) is the most speculative method advocated in the submitted articles and the one whose potential strengths remain more theoretical than empirically proven. Nonetheless, within the constrained remit of its intended usage described in the article, as a method for associating contextual information concerning a research output with the output itself, it proved itself practical and efficient: using XSLT transformations, it readily generated METS instances containing logical structural maps populated with metadata from CERIF files in a manner akin to that proposed by Habing and Cole (2008) for implementing OAI-ORE within METS. These instances could then generate contextual metadata for their respective objects in any required format, including CERIF4REF.

The potential strengths of this method are, however, more significant than the relatively narrow context within which it was first devised. Its most salient strength is to represent a way to obviate one of the most significant limitations of the intermediary schemas of Method 1, the likely proliferation of multiple bespoke schemas and of incompatible systems designed to accommodate them. It may achieve this as, like an intermediary schema of this type, it offers a means of mediating complex metadata structures such as CERIF, but does so by utilizing a constrained structure embedded within an established standard such as METS.

This method offers, with some limitations discussed below, a means of fulfilling substantially the same functions as a mediator such as CERIF4REF or CCS: it encodes a constrained version of a complex and flexible schema that has potential problems of syntactic or semantic interoperability. The strength that this method offers over more conventional intermediary schemas is that these functions are realized within a single, relatively rigid architecture, that of the METS schema itself. By using a single framework for the intermediary schema, it obviates the need for the potentially disordered environment of multiple schemas discussed above.

This should allow systems to be designed on the basis of this single schema alone: they should be more readily able to communicate with each other than those which employ diverse local schemas. A METS implementation of this type could, for instance, readily be employed as the basis of middleware such as the R4R plugins that employ CERIF4REF. This approach also has the strength of greater archival robustness than Method 1 as the METS standard is fully compatible with the OAIS reference model and METS instances are designed to act as SIPs, AIPs and DIPs within OAIS-conformant archives.

This is a significant strength of the method but a number of limitations should be noted. Although the use of the METS structural map may be a strength in terms of interoperability, it inevitably imposes limitations on the architecture of an intermediary schema of this type. The structures encoded within such a schema must follow the constraints of the striped syntax and the hierarchies of the structural map itself (possibly supplemented by structural links that are also available within the METS architecture). There will therefore be considerably less flexibility in the design possibilities of these as compared to schemas such as CERIF4REF, although the potential enhancement of interoperability discussed above may well be considered adequate compensation for this limitation.

A further pertinent issue, already discussed above in relation to intermediary schemas such as CERIF4REF, is that of synchronization; in addition to possible problems of maintaining this between a serializing structural map and its associated XSLT stylesheets, it may also be difficult to keep the semantic structures encoded within the structural map synchronized with the referent schema. Any change in the referent schema would break its link to its associated mapping in the METS structural map. Further work would be required on establishing techniques for detecting such broken linkages before this method could be employed in a working environment.

Method 3

The principal strength of Method 3, the use of the METS structural map for templating or conceptual modelling, is the extent to which it emulates many of the features of RDF-based models for managing the metadata of complex objects (particularly the Fedora CMA (Lagoze et al., 2005)) within the framework of an established XML

schema. Those features for which it does so most successfully are the flexible modelling of complex objects, enabling the reuse of components and the facilitation of processing at multiple levels of granularity.

The varying requirements of the BRIL project for the delivery of still and video images demonstrate this method's ability to model complex objects flexibly. Components may, for instance, be combined in parallel or sequentially using the <par> and <seq> elements within the METS structural map and cross-cutting structural links (using the <structLink> section>) may readily be employed. Within the apparently rigid confines of the structural map, there is considerable scope for encoding the complex metadata structures required for the delivery of compound digital objects.

This method also enhances possibilities for the flexible reuse of the components of compound objects by separating the structures of their metadata from its item-level constituents. By the use of placeholder elements within the METS file section, which are populated on delivery by separate skeletal METS files for the components from which the metadata of a complex object is aggregated, the latter can be reused as freely as required. Multiple levels of granularity, from the most abstract concepts (for instance, the data types of Article 4 (p.142)) to the smallest data components (the raw images of Article 2), can be encoded within the same integrated architecture.

There are significant advantages to enabling these features within the architecture of the METS schema. Metadata packaged in this way is easier to manage than a collection of RDF triples and is also, as noted above, compliant with OAIS for archival purposes. Despite some problems of METS interoperability which have been elaborated in the literature (for instance by Guenther (2008) or McDonough (2006)), a tightly-designed METS implementation should be more readily exchangeable and transferrable than an equivalent Fedora Content Model (as Sharma (2007) shows). One of the principle

strengths of this method, therefore, must be its ability to emulate some of the core features of the Fedora CMA within the more rigid METS framework.

Significant limitations to this method must also be noted. On a practical level, one problematic feature is that it requires a two-stage process to generate a digital object, the first to generate the final METS file in which all placeholders are populated with their respective metadata, the second to realize the digital object itself. This two-stage procedure may impact performance, although, as is noted in Article 2 (p.110), an efficient system such as Fedora can effect XSLT transformations very speedily and in practice little degradation in performance is noticeable.

More fundamental limitations arise from the use of XML architectures to express relationships between components. The METS structural map records syntactic (structural) linkages between components and not the semantic linkages which the RDF-based Fedora CMA model can more readily encompass. In many cases this is not problematic: complex objects in the latter often employ semantic links which are essentially syntactic (structural) and so can readily be incorporated within the structural map. 'Is Part Of' or 'Is Constituent Of', for instance, are two often-used relations defined within the Fedora RELS-EXT ontology (Fedora Commons, n.d.) which can be handled without problem; this is possible because the hierarchical architecture of the structural map readily models the part-to-whole associations that they represent. In other cases, however, useful relationships such as 'Has Equivalent' or 'Has Annotation' do not so clearly represent intersections between levels of a hierarchy and so are more difficult to express in this way.

At present this method is limited to the former more syntactic structural linkages and so is less rich than the Fedora ontology. It does attempt (in Article 4) to enable the Fedora 'Has Equivalent' link by employing the relatively inelegant approach of mapping to

MADS authority files (p. 145), but this a complex solution which must run in parallel to the METS environment in which more hierarchical relations are expressed and is not scalable to encompass other useful relations; it is, therefore, less satisfactory than a fully integrated set of semantic linkages within a single METS framework. Enhancing the semantic richness of potential linkages within these templates requires future work: some possible approaches to this are discussed in Chapter 6.

Conclusion

Within the stated remits of the research under which they were first devised, the three methods proposed in the submitted articles have proved relatively successful and demonstrated a number of strengths.

The intermediary schemas of Method 1 (CERIF4REF and CCS) have shown that they can enhance interoperability: CERIF4REF has proved compatible with metadata from a heterogeneous set of sources and has been tested in practical environments, while CCS, although not yet implemented in working systems, offers the potential to render the archival record more data-centric and interoperable than EAD.

Method 2 has shown itself capable of generating contextual information from complex CERIF environments and offers the potential for providing the enhanced interoperability of intermediary schemas within a constrained and standardized framework. Method 3 has proved itself able to emulate the flexible modelling, reuse of components and multiple levels of granularity of RDF-based systems for constructing aggregated complex objects.

All of these methods have their limitations. Method 1 could possibly reduce interoperability by generating a proliferation of mutually-incompatible bespoke schemas: it also suffers from a more limited validation capacity compared to architectural processing and a potential for intermediary XML schemas and their associated XSLT transformations to become unsynchronized. Method 2 could be over-constrained by the architectures of the structural map and also has the potential for its XML and XSLT constituents to become unsynchronized. Method 3 is limited by requiring two-stage processing and, more significantly, from relying primarily on syntactic and not semantic modelling for the templates that it encodes.

Despite these limitations, these methods present significant possibilities for enhancing interoperability and allowing the flexible modelling of complex objects within XML architectures. Their relevance for contemporary research is discussed in the next chapter.

Chapter 5

Contemporary relevance of this research

Introduction

This chapter attempts to address the contemporary relevance of the work documented in the submitted articles and to assess how the methods described may have a bearing on current research. It discusses three areas to which these methods are particularly relevant: these are the concept of the digital ecosystem as applied to research infrastructures, new developments in archival description, and investigations into the problematic features of RDF and Linked Open Data (LOD) in the context of digital asset management and preservation.

Digital ecosystems

The digital ecosystem has been a widely-employed metaphor over the last decade for a particular model of information and research environment. Boley and Chang in a heavily cited article (143 citations on Google Scholar) characterize such ecosystems as open communities without centralized control, in which 'swarms of agents' resolve problems as a collective, only resorting to leadership as required by the contingencies of a given problem (Boley and Chang, 2007, pp.1–4). Others develop the metaphor further by emphasizing such an ecosystem's cyclical nature (Pollock, 2011), its self-regulating system of balance and continuous development (Hitruhhina, 2012) and its self-organization and scalability (Briscoe and De Wilde, 2006, p.17).

The digital ecosystem metaphor has been applied particularly to describe collaborative research environments: Pournaras and Miah, for instance, point out the applicability of its biological metaphor for the modelling of collaborative co-ordination in these (Pournaras and Miah, 2012, p.5). In the domain of digital humanities in particular, research infrastructures or collaborative research environments have frequently been described in terms of this metaphor. Anderson, for instance, emphasizes 'change, collaboration, and engagement' as the core characteristics of such an infrastructure (Anderson, 2013, p.20) while Anderson and Blanke, when discussing such projects as DARIAH (DARIAH Project, 2017) and EHRI (EHRI Project, 2017), emphasize their derivation from communities of researchers rather than the technologies that they employ (Anderson and Blanke, 2012, p.161).

One feature of many discussions of humanities research environments is a denigration of the use of standards of all types, including metadata. Conformance to these is often argued to be obstructive to innovative research: Van Zundert, for instance, states that 'the exact purpose and need of explorative research is to go beyond what is within the standard' and so 'if a generalized digital infrastructure is to serve any meaningful humanities research it cannot be entirely governed by standards…standards run the risk of impeding rather than leveraging innovation and research' (Van Zundert, 2012, pp.173–174). Although his comments relate particularly to the humanities, they may be equally applied to any domain in which metadata, and the standards within which it is codified, forms part of the research infrastructure.

Despite the possibility of research becoming ossified in this way, abandoning standards entirely risks restricted interoperability, isolation from other research infrastructures and the danger of reinventing techniques and strategies which have already proved viable. Intermediary schemas, particularly those of Method 1, may have a role in providing a

link between the constantly-developing terrain of a digital ecosystem and the stable standards on which most metadata environments are constructed.

In a conference paper by myself and Hedges from 2013, it is argued that intermediary schemas such as the CENDARI project's CCS can form the centres of research infrastructures which can change as research proceeds to reflect its evolving requirements and those of the communities that produce it (Gartner and Hedges, 2013, p.61). The paper proposes the construction of separate schemas for each domain within an infrastructure (for instance, the medieval and early World War I areas of interest within CENDARI), which can then be supplemented by ontologies that change in response to new research as it interacts with archival materials (for instance, by annotation) (Gartner and Hedges, 2013, pp.63–64). Figure 1 illustrates the components of this proposed ecosystem model.

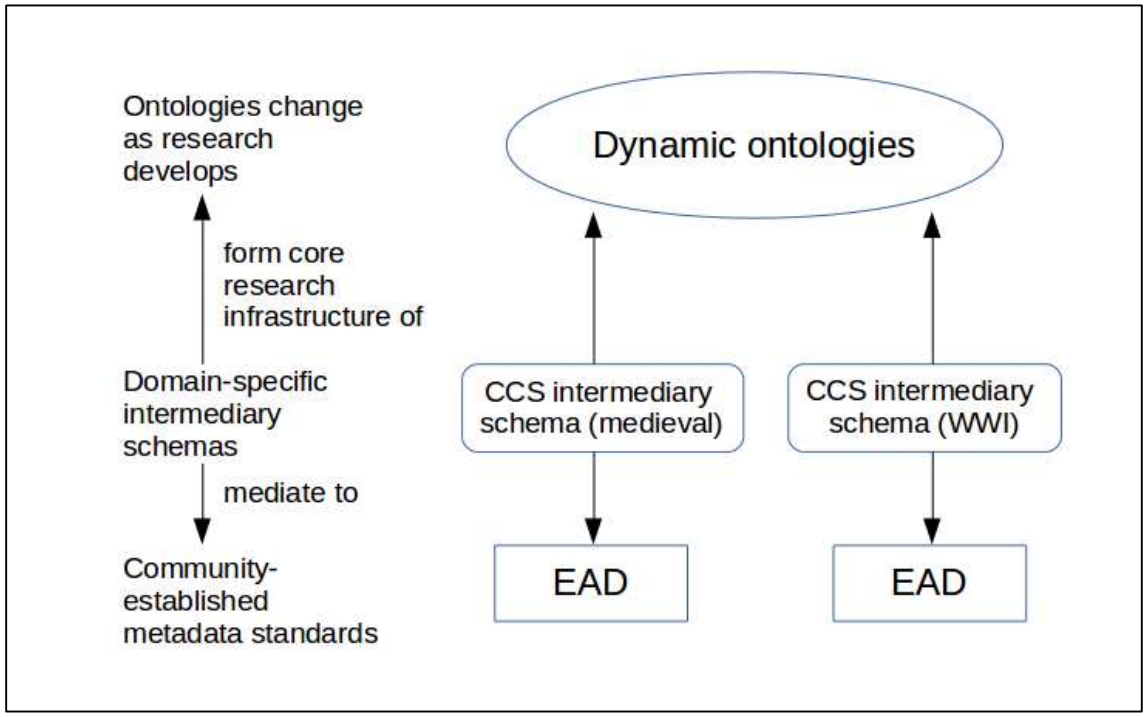


Figure 1 Components of proposed digital ecosystem model for CENDARI

These ideas were not taken forward within the CENDARI project itself because of its eventual choice of ICA-ATOM for a delivery mechanism but they offer the potential for

contributing to current research on the implementation of research infrastructures as digital ecosystems. It should be possible to use intermediary schemas in their role as constraining mediators as the core of research infrastructures, their links to schemas such as EAD ensuring that they do not become isolated from established practices and standards within their domains. Supplementing this core with more flexible ontology-based architectures, which can change rapidly as research develops, will allow the formation of digital ecosystems that are responsive to complex and changing environments. In this way, there is the potential for a standards-based infrastructure to be reconciled with the evolving requirements of one which facilitates rather than impedes research.

Current developments in archival description

A number of initiatives are currently in progress to establish new models of archival description with the aim of mitigating the problematic features of EAD discussed earlier in this critical summary. These include the design of new database architectures based on NoSQL principles and approaches which employ LOD and newly-designed ontologies. The intermediary schemas of Method 1, particularly the CCS schema, are highly relevant to these strands of research as they may realize the improvements to description that these new methods attempt to achieve without breaking entirely links to the widely-implemented standard of EAD.

The first strand of this research is the work of EHRI, the European Holocaust Research Infrastructure, which employs graph (NoSQL) databases of the type common in social media sites (Blanke and Kristel, 2013). The project's investigators suggest that the network-like structures of a graph database are effective at emulating the hierarchies of

a standard archival description but acknowledge them to be less well suited as a means of addressing the requirements of an archive for the stable and long-term persistence of research data (Blanke et al., 2017, pp.19–20).

The second strand, currently particularly active in Italy, attempts to construct new ontologies for archival description in order to exploit the potential of LOD. One project, ReLoad, attempts specifically to examine 'how using Semantic Web technologies to link to common resources…would facilitate the integration of diverse archival collections in a single web of data' (Archivio Centrale dello Stato, 2017b). A new ontology, OAD (Ontology of Archival Description) (Archivio Centrale dello Stato, 2017a), has been compiled as the primary metadata architecture for this project and XSLT stylesheets have been written to transform EAD files to RDF/XML conforming to this ontology. A further ontology, SAN Ontology (ICAR - Centro MAAS, 2014), produced by the Italian Sistema Archvistico Nazionale, also attempts to incorporate the main features of archival description into a ontology framework.

All of these projects attempt to render archival descriptions more interoperable than is currently possible with EAD owing to its document-centric focus. All are problematic to some degree. The database approach of EHRI works well in allowing the integration of a large number of archival descriptions compiled according to a heterogeneous range of standards (or in some cases to no standard), but does so at the cost of embedding these in a bespoke and complex database at a remove from standard archival practices and with limited long-term preservation potential. The two archival ontologies both address higher-levels of an archival description only using broad undifferentiated categories such as *Archival History* and *Scope and Content*, and avoid finer levels of granularity and the data-like components that are a central feature of the CCS schema.

The CCS schema and its underlying intermediary schema method offer the possibility of attaining similar results to these approaches while retaining links to current archival practice as instantiated in EAD. It data-centric components allow much of the fine-grained levels of detail that the NoSql techniques of the EHRI project can achieve and should, perhaps with further development, be as compatible with as diverse a range of metadata sources as that project claims its approach to be. It should also, perhaps with some amendments to incorporate extra facets currently missing from its element set, prove capable of accommodating archival descriptions at the levels of the ontology project discussed above and be capable in addition of finer-grained levels of detail than either OAD or the SAN Ontology can achieve at present.

The CCS schema could clearly be a useful complement to these research strands if not a replacement for them. Employing it as a mediator will ensure compatibility with EAD and the many finding aids already encoded using that standard. Embedding this metadata within an XML architecture also offers the potential for more secure long-term preservation than a series of database graphs or RDF triples. Although not in itself capable of acting as an OAIS SIP, AIP or DIP, embedding a CCS instance within an OAIS-compatible architecture such as METS would enable it to form the central component of such a package. In this way, archival descriptions can become more congruent with preservation practices than these alternatives currently allow.

Linked Open Data in digital asset management and preservation

A further area in which the methodology of intermediary schemas, particularly Methods 2 and 3, is highly relevant to contemporary research is that of the management and preservation of metadata and data encoded and stored as LOD. Multiple issues have

been cited in recent and on-going research concerning the use of LOD in digital asset and library management. A report for Jisc from 2011 on its potential for libraries (Hawtin et al., 2011), for instance, noted such issues as the complexity of data modelling and ontology development (p.17), problems of data cleansing (p.17) and skills shortages (p.90) and concluded that 'we have yet to see any real examples of benefit emerging from JISC projects in this area, or elsewhere' (p.26).

While some of these issues are probably ephemeral (skills shortages, for instance), others are likely to remain problematic owing to the inherent features of LOD. Its fluid boundaries present particular challenges for preservation: OAIS, the most commonly-adopted reference model for preservation, is designed on the principle of the discrete packaging of data and metadata. The boundaries of these packages establish spheres of responsibility for preservation and their blurring renders it difficult to know where the packages begin and end and hence what should be preserved (Gartner, 2016, p.92). Rights management can also present challenges in this environment: the Jisc report noted above cites concerns over a possible loss of control of data as a major issue for libraries (Hawtin et al., 2011, p.21; Article 5, p.154). It was possibly for these reasons that early OAIS-based projects, for example CASPAR (Giaretta, 2008) and PLANETS (Farquhar and Hockx-Yu, 2008), avoided entirely addressing the issues surrounding the preservation of LOD.

Methods 2 and 3 are highly relevant to these issues as they offer mechanisms for emulating many of the features of RDF within the more rigid and bounded architectures of XML schemas. The serialized maps of Method 2 and the templates or conceptual models of Method 3 enable the containment of potentially highly complex metadata structures within discrete METS packages; this should allow the well-established mechanisms of digital asset management to operate on these as effectively as any other

cleanly delineated digital object. This should obviate many of the problems noted by Hawtin et al. (2011), such as difficulties of data cleansing and issues surrounding any potential loss of control.

These methods also offer possible mechanisms to address some of the issues of digital preservation that have proved problematic for RDF. As previously mentioned, METS is designed to act as an OAIS SIP, AIP and DIP and so the METS-encoded instances that these methods employ are compatible with this reference model. The boundaries of these instances also clarify spheres of responsibility for preservation more clearly than the indistinct fringes of RDF-encoded metadata.

They also have relevance to more recent research into digital preservation that has moved away from the OAIS model into that of the digital ecosystem. The current PERICLES project (PERICLES project, 2017), for instance, abandons the custodial model of OAIS and its reliance on discrete packaging as less useful in dynamic environments which change rapidly and where distinctions between the active and archival phases of an object's lifecycle are less clear (Waddington et al., 2016, p.54). The model adopted in this project is that of a digital ecosystem supporting a continuum where there is not necessarily a final state for an object (such as software-based art) (Lagos et al., 2015, p.18).

The serializing and templating techniques of Methods 2 and 3 offer ways to accommodate dynamic environments without abandoning OAIS entirely. It may be possible, for instance, to combine OAIS-compatible METS templating with more flexible ontologies in the manner described in Gartner and Hedges (2013). It may also be possible to accommodate a constantly changing environment by employing METS files that encode templates or conceptual models as the core of an OAIS repository and associated METS files, perhaps incorporating maps of the type generated by Method 2,

which change to record the content of fluctuating object instances. These possibilities will require extensive further research, but they offer the potential for reconciling the exigencies of dynamic environments with the stability of established digital preservation practices in any domain.

Conclusion

The methodology of intermediary XML schemas has clear relevance to contemporary research in digital ecosystems, archival description and digital asset management and preservation. In the design of digital ecosystems, it has the potential to incorporate standards-based cores around which these dynamic environments can be constructed. In the area of archival description, it offers a data-centric approach which is nevertheless compatible with the well-established EAD schema. In the field of digital asset management and preservation, it can enable some compatibility between LOD architectures and the package-based model of OAIS.

In all three areas, it allows a degree of reconciliation between new developments and established practices and standards. In particular, it offers the possibility for flexible and often rapidly changing environments to be centred on a more stable core that can mediate to established metadata standards, so enhancing the potential interoperability of systems built on these new practices.

Chapter 6

Directions and implications for future research

Introduction

This chapter attempts to assess how the research described in the submitted articles may be taken forward in the near future. The areas of current research to which it is relevant that were discussed in the previous chapter, digital ecosystems, archival description and LOD in the context of digital asset management and preservation, will undoubtedly be developed further in future research and intermediary schemas will remain relevant to them. This chapter concentrates not on these areas but on how the methods themselves may be developed further, their implications for digital preservation in general, and the possibility of their synthesis into a more powerful methodology.

Enhancing the methods: Method 1

Method 1 has, as discussed earlier, already proved itself a viable option for enhancing interoperability between heterogeneous sources in working environments. Two limitations with the method itself, its inability to allow the validation of an instance against both the intermediary and referent schemas as part of processing and the possibilities of lapses of synchronization between a schema instance and the XSLT needed to realize it (Lubell, 2001, p.409), have already been noted.

The first issue cannot be resolved as an intermediary schema of this type is often more complex than a one-to-one mapping of elements and so the standard mechanisms of

architectural processing (specifically the definition of a FORM attribute to map elements) are of limited use. Some future research on the second issue may devise more formalized methods for associating a schema and XSLT than are currently available. Employing the <behaviorSec> within METS, as described in Article 4 (p.142), could possibly be developed further, or other ways of associating the two more intimately could be examined.

The final issue raised earlier, that a proliferation of bespoke schemas of this type may impede interoperability in general, raises questions about whether developing intermediary schemas of this type should be followed up in future research. The discussion below, on synthesizing the other two methods discussed in these articles, suggests an alternative to this method which could resolve these issues.

Enhancing the methods: Method 2

This method is the most speculative and so requires the most development in future research to enable it to function as a reliable and robust technique in working systems. It could, however, as argued below, form the basis of a synthesis of the three methods with significant implications for future developments.

Future research will be needed to resolve the limitations of the method that were highlighted in Chapter 4. The constraints of the METS structural map and the limitations of the striped syntax which this method employs need alleviating to allow it to capture the full complexity of all potential objects and some research will have to focus on this. In addition, the method currently extracts semantics only from its referent metadata, defining the maps it creates by the selection of these components but not

adding additional layers of semantics which could define potentially richer models. Some mechanism for incorporating additional semantics, for instance by integration with external ontologies, should be researched to enhance the method in this way.

The most significant issue with this method as it is currently conceived is its vulnerability to any changes in the schema to which the instances that it maps conform. Any future development will require the embedding of checking mechanisms to ensure the maps that it encodes are synchronized with the schema. This would best be achieved by devising a method for defining an abstraction of the map itself which could be compared with the referent schema; any changes to the referent should become apparent as anomalies in the process of comparison. Research would be required to identify a suitable method for encoding this abstraction and formulating the techniques by which it could be compared to the referent schema.

Enhancing the methods: Method 3

A significant issue that has already been highlighted with Method 3 is that the linkages that can be expressed in the METS structural map are syntactic: they express structural relationships such as ‘Is Part Of’ instead of semantic relationships such as ‘Has Equivalent’. This is a significant impediment to expressing relationships of the richness that is possible within RDF-based models such as the Fedora CMA. Further development is required to allow semantic relationships of this type to be included in these templates.

A number of possible solutions to this problem already exist. A basic but crude one would be to employ the METS LABEL attribute which is available within elements to express these relations: this is the technique deployed by Habing and Cole (2008) in their striped syntax to express OAI-ORE relationships within the structural map. Employing textual labels in this way, however, is very imprecise and potentially error-prone. A more sophisticated approach would be to employ METS's structural linking facilities: the <structLink> element may contain an XLink ARCROLE attribute which can contain a URI expressing a semantic relationship. This could then be used to express linkages of this kind between elements in the structural map. Other possibilities, some of which may be facilitated by moves towards revising METS into a data model more congruent with the Semantic Web (METS Editorial Board, 2010; METS Editorial Board, 2011), should also be explored in future research.

Enhancing digital preservation

One area of future research to which intermediary schemas may make a significant contribution is that of digital preservation. The serialized maps of Method 2, in particular, may be relevant to future developments in preservation metadata. It would, for instance, prove fruitful to examine how they might be used in conjunction with PREMIS metadata (Library of Congress, 2015) within a METS framework as a way of defining, or at least presenting, views which could clarify intellectual entities for digital objects.

In addition to PREMIS, a further promising direction for future research is likely to be in the context of the OAIS reference model. As noted earlier, METS is designed to act as an SIP, AIP and DIP within the context of OAIS and so the METS architectures

within which this proposed synthesis is situated ensure some degree of compatibility with this framework.

A specific model which should be researched is one in which instance data is packaged with the serialized maps of Method 2; these could act as representation information on the archived data, facilitating its decoding when disseminated as a DIP. These maps could be tailored to specific Designated Communities, 'the set of Consumers who should be able to understand the preserved information', as defined by the OAIS model (Consultative Committee for Space Data Systems, 2012, pp.2–3). A formalized method for encoding representation information in such a manner could prove a valuable area of future research within the digital preservation community, particularly within the GLAM sector.

Synthesizing the methods

The final, and potentially most productive, line of future research into developing this methodology further would be an investigation into how its subsidiary methods could be synthesized: more specifically, this would be an analysis of how the templating and conceptual modelling of Method 3 could be integrated with the serialization of metadata structures of Method 2 in order to create a new method which would fulfil most of the functions of the constraining mediators of Method 1. The benefits accruing from such an approach would be to produce a more standardized technique for defining constrained intermediaries to referents without the proliferation of numerous bespoke schemas which could impede interoperability overall.

Both templating and serialization act as maps for instances of complex applications such as CERIF, although they operate in opposing directions: templating is a mapping

which generates instances while serialization produces maps of those instances once generated. In combination, the two methods have the potential to act as mediators to established schemas as follows:-

- The maps produced by Method 2 could be used to facilitate the ingest of data and metadata from diverse sources (as has been achieved, for instance, by the CERIF4REF schema)
- A higher level of mapping could then consolidate these multiple maps into a more definitive specification of the metadata architecture of the domain within which these sources are located; this could be an abstraction of the type discussed as an enhancement to Method 2 above
- If required by the applications which will process and deliver the metadata that is mapped in this way, formal XML schemas could be generated from this higher-level mapping
- Templates or conceptual models (Method 3) could be defined from this higher-level mapping: using these, instances conforming to the referent schemas can be generated.

Figure 2 (overleaf) illustrates diagrammatically this proposed synthesis.

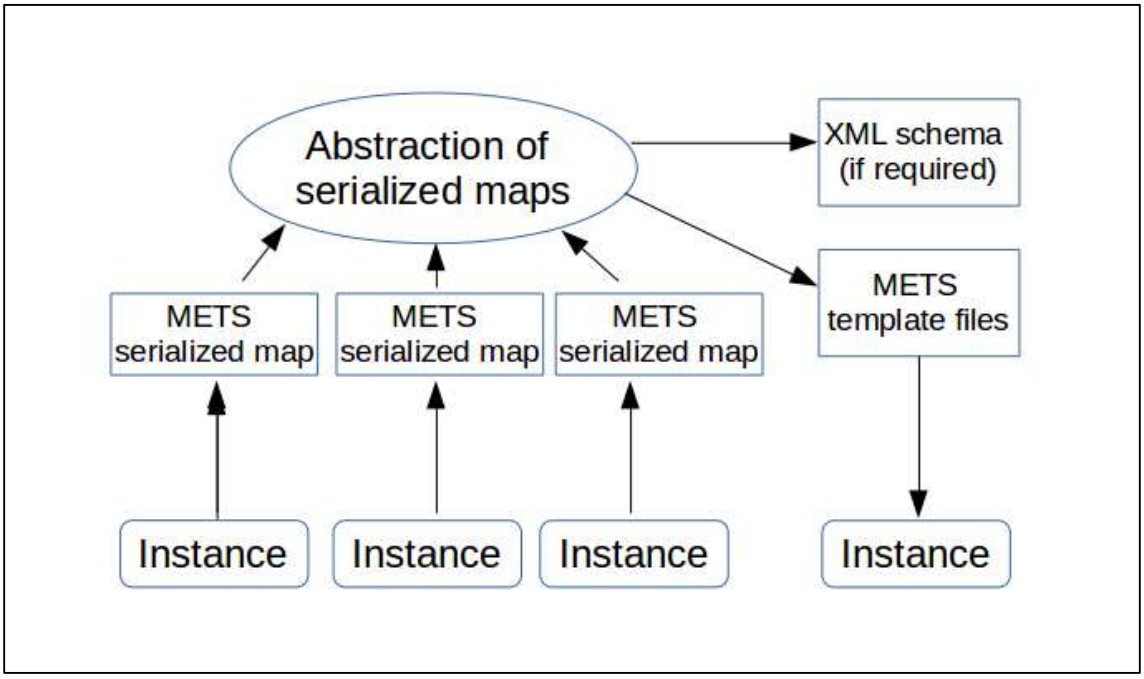


Figure 2 Model for proposed synthesis of three intermediary schema methods

To achieve such a synthesis would require an extensive programme of research of at least the length and complexity as that which developed the methods documented here. It does, however, offer the potential for standardizing the intermediary schema methodology within the structures of the established METS standard and avoiding the heterogeneity of an environment in which separate schemas are constructed for each of a range of diverse applications. In doing so, it would increase substantially the applicability of this methodology to working environments.

Conclusion

The three methods that comprise the methodology of intermediary schemas can form the basis of significant and fruitful future research. Each can usefully be developed further to overcome such current impediments as issues of synchronization between

intermediary schema instances and the XSLT transformations that realize them (Method 1), the constraints of the striped syntax used to serialize metadata structures in the METS structural map (Method 2) and the limited semantic linking facilities of this map for templating (Method 3). Method 2 in particular also offers the possibility for future research into OAIS-centred digital preservation as an aid to the dissemination of archival objects to Designated Communities.

The most productive pathway for future research would be to synthesize these methods within an overall METS framework as described above. This has the potential to formalize the overall methodology within a single architecture, avoiding the possibility of a disordered environment in which multiple intermediary schemas propagate and interoperability becomes more difficult to achieve. METS is now a well-established, community-embedded standard that is most likely to remain widely adopted in the long term and so offers a dependable environment within which this methodology can function efficiently.

Chapter 7

Conclusion

The articles submitted for this PhD by Prior Publication and this accompanying critical summary introduce the methodology of intermediary XML schemas and its three subsidiary methods which, in differing ways, apply its core concept of schemas which mediate either to others or to alternative instances of metadata conforming to themselves. Method 1 employs project-specific intermediary schemas which act as more constrained and hence more interoperable mediators to their referent schemas. Method 2 employs the established METS schema as a way of mediating to a referent schema by serializing a map by which the latter's complex instances may be navigated. Method 3 uses the same METS schema as a means of encoding templates or conceptual maps of the metadata for complex compound digital objects which are populated on delivery by XSLT transformations.

All three methods extend significantly earlier approaches taken to enhancing interoperability at a schema level, such as crosswalks, derived schemas, switching across or application profiles. Method 1, although bearing some relation to crosswalks or application profiles, extends these techniques to incorporate more fluid and complex relationships than these antecedents are capable of assimilating. Method 2 employs techniques which can be applied to aggregating compound objects, such as OAI-ORE and the METS structural map, but extends them to treat complex metadata as such an object. Method 3 represents an advance on earlier projects which sought to use XML schemas for templating by obviating the need for dedicated middleware or changes to the XML schema language and by allowing more sophisticated modelling than earlier METS-based methods have achieved.

All three methods have demonstrated their viability in the context of the projects in which they originated. CERIF4REF, the Method 1 intermediary schema documented in Article 1, proved capable of mediating between heterogeneous metadata sources and formed the basis of the development of software plugins to several repository systems. Method 2 proved itself viable as a means of associating contextual information with a research output and generating CERIF4REF output. Method 3 proved its potential in the BRIL project as a means of generating complex compound objects in a manner analogous to RDF-based Fedora content models within an XML environment.

The methods have particular relevance to three areas of research which are currently being actively pursued. The application of digital ecosystems to research infrastructures may be enhanced by the use of the intermediary schemas of Method 1 as their core, supplemented by ontologies which develop to reflect dynamic environments that change as research develops: this can ensure flexible research environments which are nevertheless linked to established standards and practices. Current research into archival description which is concerned with enabling interoperability between heterogeneous sources may also benefit from the same method: it should be capable of integrating diverse sources at fine levels of granularity as effectively as the database- or ontology-based approaches being investigated at present. Methods 2 and 3 are highly relevant to contemporary digital asset management and preservation research as they can enable the features of RDF-based approaches to these to be contained within OAIS-compliant METS frameworks.

All three methods can be developed further in future research. Method 1 may benefit from a more reliable means of ensuring synchronization between a schema and the XSLT used to transform its instances. In view, however, of its potential for impeding interoperability by producing a proliferation of schemas, it may be superseded by developments in the other two methods introduced here. Method 2 may benefit from

research into extending the striped syntax used for serialization, examining new ways of integrating its maps with external ontologies and identifying more dependable methods for ensuring that its mappings remain synchronized with the referent schemas whose instances it serializes. Method 3 could benefit from further research into enhancing the semantic scope of the relationships encoded in its templates, extending the primarily syntactic character that they currently have.

The most fruitful line of future research based on this methodology, however, would be to attempt a synthesis of Methods 2 and 3, an approach which would provide the advantages of interoperability provided by the constrained mediation of Method 1 but without the necessity of designing project-specific schemas which threaten to undermine it at a macro level. Such a synthesis would employ the serialization techniques of Method 2 to create maps from which the templates or conceptual models of Method 3 could be defined. Such a synthesis could be particularly useful in the context of preservation metadata as it could provide a basis for standardizing representation information within the context of OAIS.

Research questions

This section of the conclusion attempts to address and answer the two research questions posited in the introduction as they relate to the methodology of intermediary schemas as a whole.

Research question 1: in what ways and to what extent may intermediary XML schemas be employed to facilitate the management, preservation and dissemination of metadata for complex digital objects?

All three methods may prove useful in the management, preservation and dissemination of the metadata for complex objects. Method 1 can facilitate these by constraining and simplifying the complexity of this metadata through the definition of schemas to support instances that are clearer, better structured and more constrained that those conforming to a referent schema. Method 2 can achieve a similar end by formalizing maps which fulfil comparable functions within the standardized framework of a METS structural map. Method 3 can prove of great value as a means of formalizing templates for the metadata of complex objects which can incorporate the flexibility of RDF-based architectures into an XML environment.

As indicated above, Method 1 has proved itself a viable aid to metadata management by supporting plugins for the R4R project. The constrained and simplified CERIF4REF schema facilitated the design of these more readily than the complex CERIF objects to which it mediated. CCS, the Method 1 intermediary schema introduced in Article 5, offers the potential for enabling new approaches to the dissemination of archival finding aids, such as dynamic research guides, owing to its more data-centric approach than that adopted by its document-centric referent schema EAD. The value of this method may be tempered, however, by the possibility of it producing a messy overall environment populated with multiple autonomous schemas which would reduce interoperability overall.

Method 2, although at present highly speculative, offers the possibility of facilitating the management of complex digital objects in a more formalized and systematic framework than Method 1. The maps serialized within the METS structural map by this method can act as navigation aids for complex metadata architectures which could simplify the processes of managing the complex objects that they document. Similarly, they may facilitate their preservation and dissemination by virtue of acting as representation

information, potentially offering a standardized method for encoding this information within OAIS.

Although potentially a more viable method than Method 1 because of its capacity to offer a more standardized overall architecture of greater internal interoperability, Method 2 requires the most research and development of any of these three to enable it to function effectively within working environments. It is less flexible than Method 1 as its maps must be congruent with the relatively rigid METS architecture; in addition, its striped syntax as employed in Article 3, which emulates the structure of RDF triples, is relatively basic. Both of these issues should be addressed to render the mapping facilities of the method more sophisticated, although this should not be at the expense of over-flexibility and hence lower interoperability.

Method 3 can facilitate the management, preservation and dissemination of the metadata for complex objects by virtue of its ability to emulate the flexible modelling and reusable content features of RDF-based architectures, most notably the Fedora CMA. Articles 2 and 4 demonstrate how it may be employed to manage metadata for complex objects within a single XML architecture. Separating the structure of this architecture from object-level metadata allows great flexibility in the reuse of components, comparable to that of the Fedora CMA or other RDF-based architectures; it also avoids the serious issues of overall management, preservation and intellectual property that are often associated with LOD in digital asset management. At present this method is hampered slightly by the relatively limited range of syntactic relations that can be expressed within the METS structural map, although this may be alleviated by some of the approaches discussed in the preceding chapter.

Research question 2: How may intermediary schemas enhance metadata interoperability within XML architectures?

The methodology of intermediary schemas as a whole can enhance metadata interoperability by introducing constraints into schemas which suffer from what McDonough (2008) terms excessive 'flexibility, extensibility, modularity and use of abstraction'. Method 1 does so by designing heavily constrained schemas which can nevertheless mediate to their problematic referent counterparts by XSLT transformations; this enables greater interoperability without divorcing an application entirely from practices which are embedded within their respective user communities, as instantiated in schemas such as CERIF or EAD.

Method 1 may, however, be impeded in enabling greater interoperability by the possibility, noted above, of it producing a proliferation of mutually-exclusive bespoke schemas. It may also prove problematic to ensure an intermediary schema remains synchronized with the XSLT stylesheets required to translate its instances to those conforming to its referent schema: if this synchronization is broken, interoperability with its referent schema is also lost.

Method 2 is also a method whose primary rationale is to enhance metadata interoperability. This approach aims to introduce constraint to a complex metadata environment by creating a map by which it is abstracted and serialized: this offers the constraint of Method 1 not by creating a new schema but by its selection of components and the definition of semantic relationships between them. This selected, limited and circumscribed map of components and relationships offers the possibility of greater interoperability than a complex metadata environment in which McDonough's 'flexibility, extensibility, modularity and use of abstraction' may predominate. Such maps may fulfil some of the same interoperability-enabling functions as crosswalks or switching across but offer more sophisticated features, including more complex semantic relationships, than is possible using these established techniques for schema-level interoperability.

Method 3 is primarily designed as a mechanism for enhancing the flexible delivery of complex objects by separating the structures of their metadata from its content and enabling their combination on the fly for dissemination. It also has the potential to enhance interoperability at the level of the templates which it encodes within the METS structural map. These are constrained representations of the structure of complex objects, which, like the maps of Method 2, are more readily interchangeable between systems than the complex objects that they model.

Both research questions are addressed and answered in part at least by the proposal for synthesizing Methods 2 and 3 advocated in the preceding chapter. Such a synthesis would combine the enhancement to management, preservation and dissemination offered by the flexible modelling and reuse of components of Method 3 with the enhanced interoperability offered by the navigational mappings of Method 2. For this reason, a synthesis of these two methods offers the greatest potential for realizing the potential of this methodology.

The ways in which intermediary schemas may facilitate the management, preservation and dissemination of metadata for complex objects, and their potential for enhancing the interoperability of this metadata, demonstrate clearly how they may achieve a degree of reconciliation between the antithetical requirements of flexibility and interoperability within XML architectures: this is the most significant contribution to knowledge of this research. They also demonstrate how new developments (particularly rapidly changing environments) and existing practices and standards in metadata may be reconciled so that these developments do not abandon established, community-based principles. By allowing the benefits of RDF-based approaches to complex metadata.to be emulated in XML environments they strengthen XML as a viable format for digital asset management in the long term.

Concluding remarks

The methodology of intermediary schemas introduced in this PhD by Prior Publication is an original approach to issues of interoperability and the management of complex metadata environments. Its three subsidiary methods move significantly beyond earlier approaches to enhancing interoperability at the schema level or handling the metadata for complex objects. Method 1 represents a significant advance on earlier practices such as crosswalks and Method 3 on prior attempts to use XML schemas for templating. Method 2 is the most original of all in its approach to mapping complex metadata environments in a manner akin to compound objects.

The methodology has significance for many aspects of digital asset management, preservation and dissemination, particularly if a synthesis of Methods 2 and 3, as described in Chapter 6, can be achieved. It should enable greater interoperability of metadata (and hence data) by mediating complicated environments in more sophisticated ways than earlier techniques can achieve. It allows sophisticated templating and conceptual modelling which enables the flexible reuse of metadata components in dynamic environments. It will make it possible for these to be achieved wholly within XML architectures, so alleviating some of the significant problems noted earlier in this critical summary with using RDF-based metadata for digital asset management.

Future research, as detailed in Chapter 6, will enable this methodology to be more firmly embedded within working practices but the five articles submitted here, and this critical summary, demonstrate its potential and show that the foundations to enable its potential to be realized are clearly established.

Article 1

Intermediary schemas for complex XML applications: an example from research information management

Journal of Digital Information. 12 (3), [online]. Available from: https://journals.tdl.org/jodi/Article/view/2069 (Accessed 27 February 2017)

Journal of Digital Information, Vol 12, No 3 (2011)

Intermediary schemas for complex XML applications: an example from research information management

Richard Gartner
Centre for e-Research, King's College, London
richard.gartner@kcl.ac.uk

Abstract

The complexity and flexibility of some XML schemas can make their implementation difficult in working environments. This is particularly true of CERIF, a standard for the interchange of research management information, which consists of 192 interlinked XML schemas. This article examines a possible approach of using 'intermediary' XML schemas, and associated XSLT stylesheets, to make such applications easier to employ. It specifically examines the use of an intermediary schema, CERIF4REF, which was designed to allow UK Higher Education institutions to submit data for a national periodic research assessment exercise in CERIF. The wider applicability of this methodology, particularly in relation to the METS standard, is also discussed.

1. Introduction

The complexity of important XML schemas may often present a major hurdle to their adoption, particularly in cases where they are required in environments which do not already have significant experience in XML authoring or editing. These problems may be exacerbated in cases where these schemas are highly flexible, the lack of constraint in the way in which they can be used often requiring extensive work on initial information architectural design before they are implemented in practice: the Text Encoding Initiative (TEI), for instance, usually requires a process of tag selection and semantic specification (as described, for instance, in Welty and Ide 1999, 62) unless a pre-constrained variant (such as TEI-Lite) is employed.

Such problems of complexity and over-flexibility become more acute in applications which employ multiple, linked XML files to capture the complexities of their required architectures. In such cases, not only must the elements and other components be determined, but also the form of the linkages between files (including the definition of any ontologies used to define any links with semantic meaning). The possibility of offering an 'off-the-shelf' scheme with a minimal learning curve becomes less likely the greater the complexity of the overall information environment to be encoded, and the potential take-up of any such schemas will inevitably become more limited.

This article examines one potential approach to obviating these problems in the form of 'intermediary' XML schemas and XSLT stylesheets. The context in which it is examined is that of the CERIF format (European Organisation for International Research Information 2010b), a complex application designed to facilitate the interoperability of research management information. CERIF is maintained by euroCRIS, the European Organisation for International Research Information, and offers a comprehensive model for all metadata necessary to maintain current research information systems (CRISs) in a readily interchangeable form. The CERIF model, originally instantiated as a set of relational SQL tables but since 2006 available in XML, is based on a small number of core components and an extensive set of linkages between these which can mirror their often complicated inter-relationships. Such an approach has the benefit of being able to fulfil the metadata requirements of almost any operating CRIS, but also the downside of potentially great complexity.

The methodology advocated here to alleviate some of this complexity is an intermediary XML schema and associated XSLT stylesheets which are used to select a relevant subset of the CERIF components and constrain the manner in which they are employed. This technique is employed in the context of the Readiness for REF (R4R) project (Centre for e-Research 2011) from the United Kingdom's higher education community which sought to examine the feasibility of employing CERIF in the context of the periodic research assessment exercises which are used to determine the allocation of research funding to universities and other research institutions in that country. Its aim was to render the CERIF standard, which the higher education funding body had specified as a format for submissions to the next exercise in 2014, a feasible option for the first time for the majority of institutions who had not previously found it viable.

2. CERIF and its implementation challenges

The need to share information required for research management has long been recognised, particularly where this research is publicly funded and there is a consequent onus to ensure transparency in determining the allocation of funding and ensuring that it is well spent. In addition, the international nature of much research collaboration also requires this information to be readily shared across national (and often linguistic) boundaries. Early work on rationalising the metadata necessary for sharing information of this type began in Europe in the 1980s and eventually produced the CERIF standard, which has undergone several major revisions (in 2000, 2004 and 2006) since its first appearance in 1991 (European Organisation for International Research Information 2010c).

CERIF was conceived as a data model which is independent of any given syntax, although it was initially made available as SQL tables and later as a series of XML schemas. The model defines three 'base' entities **project**, **person** and **organisation unit**, all of which include only very basic metadata (including crucially unique IDs) as shown in the diagram below:-

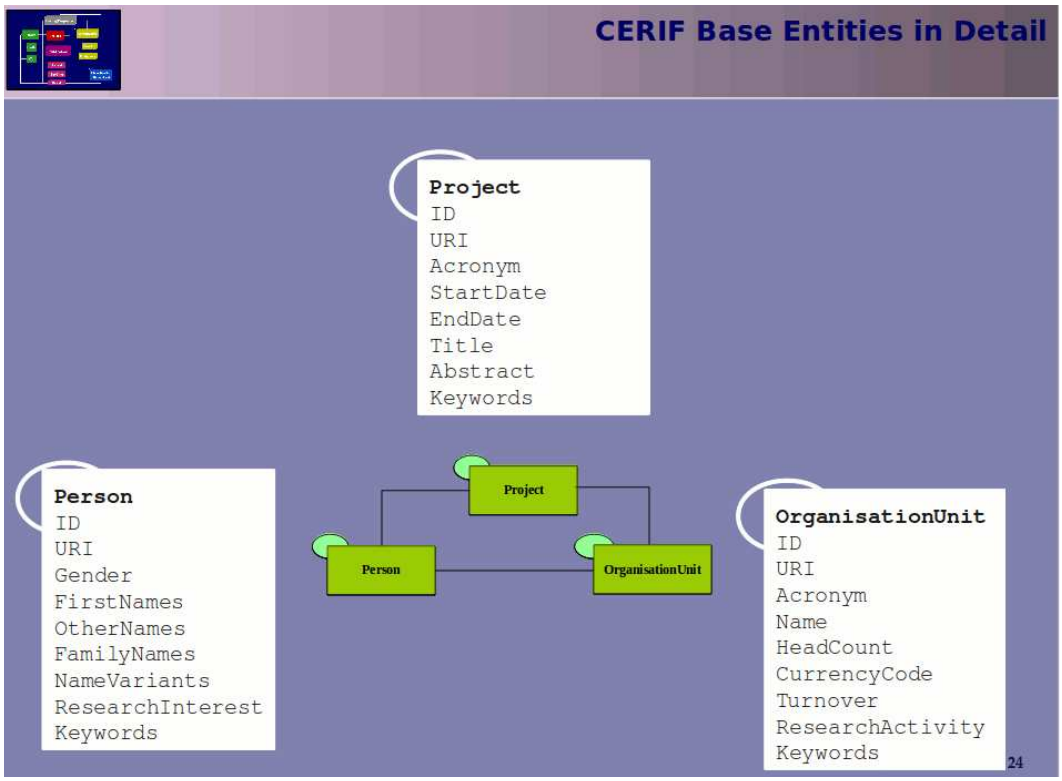


Figure 1. CERIF Base entities.

In addition, the CERIF core provides basic metadata for research outputs ('results') including publications, patents and products:-

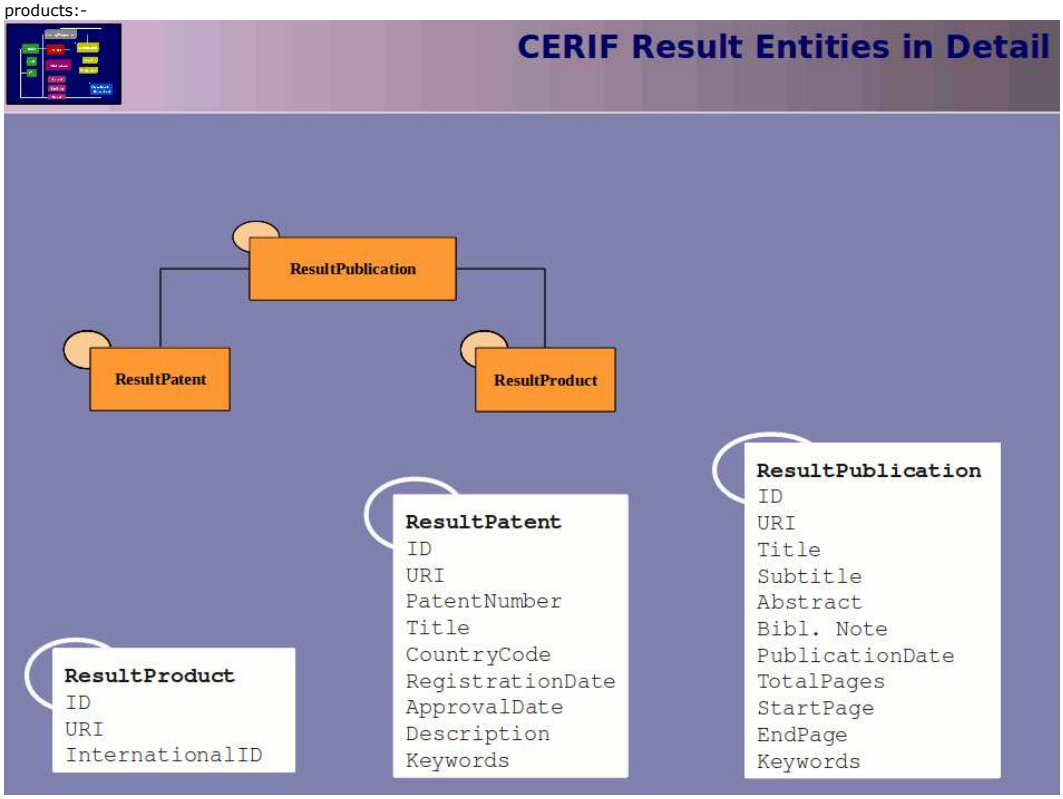


Figure 2. CERIF Result entities.

and a secondary tier of metadata components, which encode concepts potential relevance to any of these base entities such as addresses (postal or electronic), countries or events .

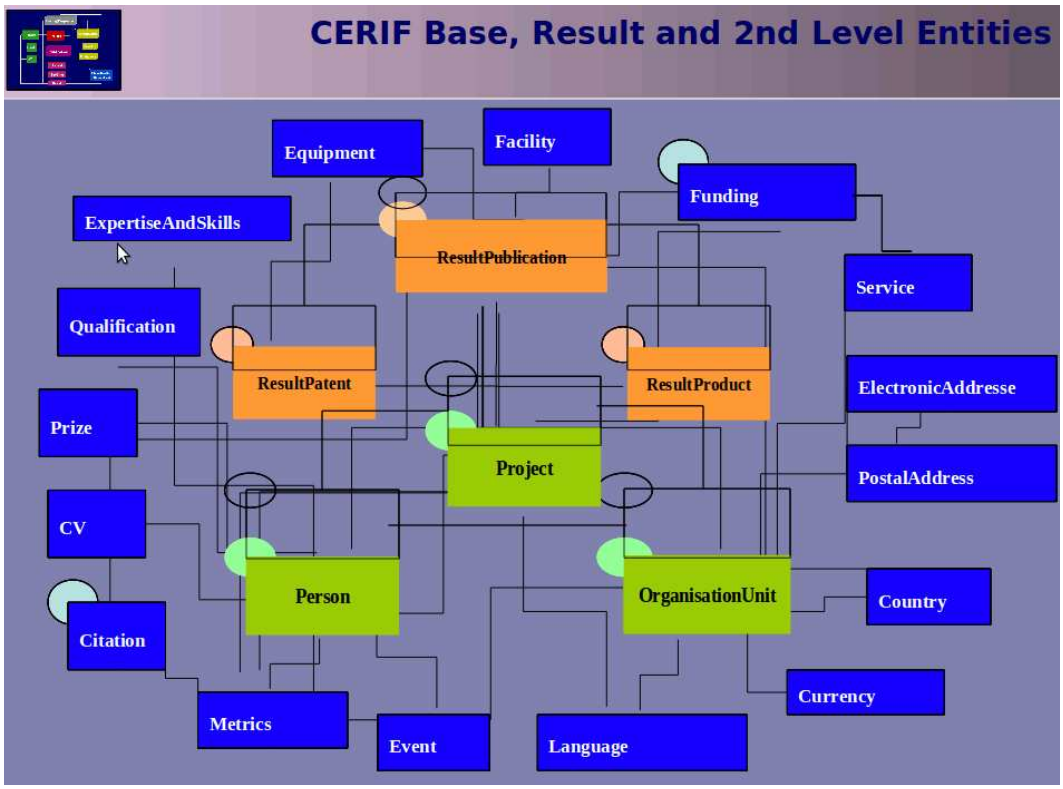


Figure 3. CERIF Base, Result and 2nd Level entities.

A final group of entities, and an additional layer of complication for the implementer, allows CERIF to function in a multi-lingual environment. Any textual data capable of translation into multiple languages, such as publication titles, abstracts, descriptions of funding programs or descriptions of research environments, must be encoded using these entities: the following fragment, for instance, contains multiple language versions (English, Italian and German) of the title of this paper:-

```
<cfResPublTitle>
    <cfResPublId>1abcdefghijklmnop</cfResPublId>
    <cfTitle cfLangCode="en-UK">Intermediary schemas for complex XML a
    <cfTitle cfLangCode="it">Schemi intermediari per complesse applica
    <cfTitle cfLangCode="de">Intermediäre Schemata für komplexe XML-Ar
</cfResPublTitle>
```

The core of any CERIF application, and the bulk of the standard itself, consists of an extensive series of linking tables or XML files which allow these base and second level entities to be joined together. These linking entities (of which there are currently 95, far more than the 3 Base and Results entities and 16 Second-level entities (Jörg et al. 2010, 36-38)) allow for a potentially complex set of relationships to be expressed which can model almost any research environment: one linkage, for example, may be used to join a member of staff to a research group, a research output, an event (such as a conference presentation), or to other researchers (for instance project collaborators). Similarly a research publication may be joined to its funding stream, to a conference at which it is presented or to a prize awarded for it.

These entities rely upon either user-defined, or preferably pre-published, semantic schemes to assign meaning to the linkages within a given application. The following XML fragment, for instance, is an example of part of the linking entity which joins research staff to their publications or other research outputs:-

```
<cfPers_ResPubl>
    <cfPersId>cerch-2008-001</cfPersId>
    <cfResPublId>publ-9999-a-0001</cfResPublId>
    <cfClassId>is-author-of</cfClassId>
    <cfClassSchemeId>cerif-person-pub-roles</cfClassSchemeId>
    <cfStartDate>2001-01-07T00:00:00-00:00</cfStartDate>
    <cfEndDate>2099-12-31T00:00:00-00:00</cfEndDate>
</cfPers_ResPubl>
```

Here the person identified by the **cPersID** element is linked to the publication (designated by the **ResPubId** element) of which he/she is the author: to designate this semantic relationship, it is necessary to declare both the semantic scheme used (in this case one designated by the identifier in **cfClassSchemeId**) and the semantic term itself (identified in **cfClassId**).

To be able to use CERIF in a real-world application, therefore, requires the identification or definition of an extensive set of semantic schemes and their consistent application. This may be particularly problematic as CERIF records lose much of their interoperability without the application of a coherent semantic scheme. Although euroCRIS have themselves published a core set of semantic terms (European Organisation for International Research Information 2010a) which would, if widely adopted, move towards resolving this problem, it at present covers only a proportion of the relationships likely to be required in a real-world application.

The complexities involved in implementing CERIF should be apparent from even this short introduction to it. This complexity is not alleviated in the XML instantiation of the CERIF data model which translates directly the structure and content of the original relational database tables that formed CERIF's first version as a SQL application. Adopting this approach was done

with good reason, particularly to retain the degree of flexibility present in the original data model which would be very difficult to replicate in a single XML schema. Using XML does result in the loss of some potentially valuable integrity rules (such as uniqueness and referential integrity) which are present in the SQL model of CERIF (Jeffery, Lopatenko & Asserson 2002, 80) , but the essential ability to model complex research environments remains intact in this use of XML. The disadvantage is the verbosity and complexity involved in employing the 192 XML schemas which form the model in this format.

3. Constraining the XML metadata universe

Much of the preceding discussion leads to the conclusion that for CERIF to achieve its potential as a medium for the interoperability of research information it requires some degree of constraint in its application. In addition, the terms under which it is constrained (for instance, the choice of schemas and the semantics to be employed) need to be adopted in an environment that extends beyond a single institution (preferably the whole research management community).

XML as a language offers fewer opportunities for constraining and validating content than are available in, for instance, a conventional relational database: the XSD schema language only allows constraint by domain range (constraining values, usually numerical, to a given range), mixed content (constraining the number and orders of child elements for complex element types) and cardinality (minimum and maximum occurrences of an element) (Jacinto et al. 2002, 29). As is well known, XML validation procedures can test the conformance of a document *syntactically* but not *semantically* (Jacinto et al. 2002, 2); this requires any desired semantic constraint to be hard-coded into a schema (for instance, in the form of an enumerated list) so that it be validated as if it were a syntactic rather than a semantic requirement. The possibilities offered by such simple validation procedures as lists or constrained attribute values are very limited, however, and the requirement for them to be incorporated into the schema when it is written makes their potential irrelevant when seeking to constrain standards such as CERIF which have already been published.

Two well-established methods for constraining the content of XML files already exist in the form of **XCSL (XML Constraint Specification Language)** (Jacinto et al. 2004) and the more widely-used **Schematron** (International Standards Organisation 2006), both of which offer the possibility of validating the content of an XML file in addition to its syntactical conformance. Both work by allowing conditions for element contents to be tested against specified contexts: for instance, the content of an ISBN element can be checked to ensure that it conforms to the required format and that it validates correctly against its check digit. Both also offer the possibility of conditional validation, so allowing, for instance, the value of a given element or attribute to dictate the structure or content of other components of a file: a poem, for instance, with a *type* attribute which can be set to such values as 'sonnet' or 'quatrain' could have its overall structure validated according to the value of this attribute (Jacinto et al. 2002, 14-20).

Neither solution is ideal for the challenges presented by a complex CERIF application. A comprehensive validation system would require 192 XCSL or Schematron files, one for each potential CERIF XML instance (although it is unlikely that any given application would in reality use anything near this potential number). The validation of CERIF's complex linkages is possible using either approach (for instance, by employing the **document()** function in XPath within a Schematron file), but rapidly becomes complicated and hard to maintain accurately once the number of linkages exceeds a relatively small number. In addition, for relatively non-technical users, the further validation steps required to use XCSL or Schematron add to the gradient of the learning curve for CERIF implementation.

4. CERIF in the context of the Research Excellence Framework (REF)

These problems became all too evident in the context of a research project undertaken at King's College London which sought to make CERIF a viable option for institutional submissions to research assessment exercises. The Readiness for REF (R4R) project (Centre for e-Research 2011) takes its name from the UK Government's Research Excellence Framework (REF) programme (Higher Education Funding Council for England 2010), the latest in a series of regular assessments of the research outputs of higher education institutions on the basis of which research funding is allocated to these bodies. It has been announced that the next exercise, due in 2014, will accept submissions in CERIF as its preferred format, although no detailed specification of the CERIF implementation envisaged has yet been published.

To test the feasibility of using CERIF as the medium for submissions, the R4R project undertook a detailed mapping to it of the data requirements from the previous exercise, undertaken in 2008 (Higher Education Funding Council for England 2008). This exercise, at that time called the Research Assessment Exercise (RAE), allowed submissions in XML which conformed to a schema devised specifically for this purpose. In the absence of any specification of the data requirements for the REF itself, this schema formed the basis of the mapping exercise. The results revealed the complexity of the task ahead, as most concepts in the RAE schema proved to mappable only by employing three or four inter-linked CERIF files: for instance, linking a researcher to the title and bibliographic details of a research publication involves a minimum of four files linked as follows:-

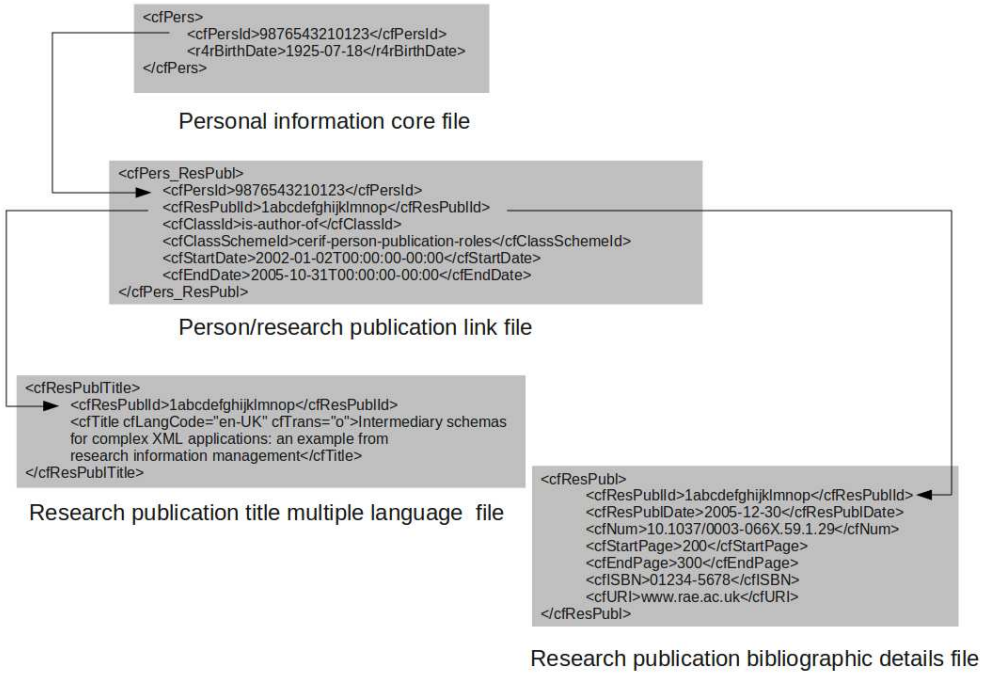


Figure 4. Researcher/research publication linkages in CERIF.

The mapping exercise concluded that in total 19 of the 192 possible CERIF files were required to encode all of the metadata specified in the RAE exercise (Gartner & Grace 2010, 100). While the number of files required was relatively small, the complexity of the linkages required and the complex semantic vocabularies required to enable them to be formed appeared daunting, and would probably have rendered the standard an inappropriate solution for institutions without an advanced technical knowledge of CERIF and XML itself.

5. Constraining CERIF with XSD and XSLT

This problem of complexity required some method of constraining CERIF when it is used for this specific application: this constraint is both syntactic (limiting the XML files used and the manner in which they are linked) and semantic (limiting the range of vocabularies used to enable the linkages). Before the advent of XML SGML provided a mechanism to allow constraints of at least the former kind to be imposed: architectural processing could be used to derive small project-specific DTDs which could then be automatically processed to larger, more complex applications constrained to the requirements of the project. Simons, for instance, uses this technique to map highly-constrained DTDs to the complex and highly flexible TEI (Text Encoding Initiative) (Simons 1998 & 1999). Such architectural processing is not available in XML but much of its functionality can be duplicated by creating highly constrained 'intermediary' XML schemas which are then processed by XSLT to created the complex application required.

The XML schema devised for this project, called CERIF4REF, is properly termed intermediary as it mediates between the requirements of the RAE schema and CERIF. It was not possible simply to devise a stylesheet to translate RAE directly into CERIF as so much of RAE consists of aggregations from a variety of data sources. In the CERIF model, these sources are encoded explicitly, and it is impossible to disaggregate them from the summary form in which they are given in RAE: for instance, the RAE requires a simple count of full-time equivalent research assistants, whereas the CERIF model requires them to be listed individually. In addition, the RAE schema makes no use of XML IDs and IDREFs for linkages, making it impossible to validate these accurately and rendering the construction of the complex network of links in CERIF much more difficult.

The CERIF4REF schema is structured in a similar way to RAE, dividing into five sections:-

1. research groups
2. research personnel
3. research outputs
4. funding
5. overall descriptions of the research environment

A comparison of the encoding of research groups and personnel in RAE and CERIF4REF provides an immediate example of the different approaches of the two schemas:-

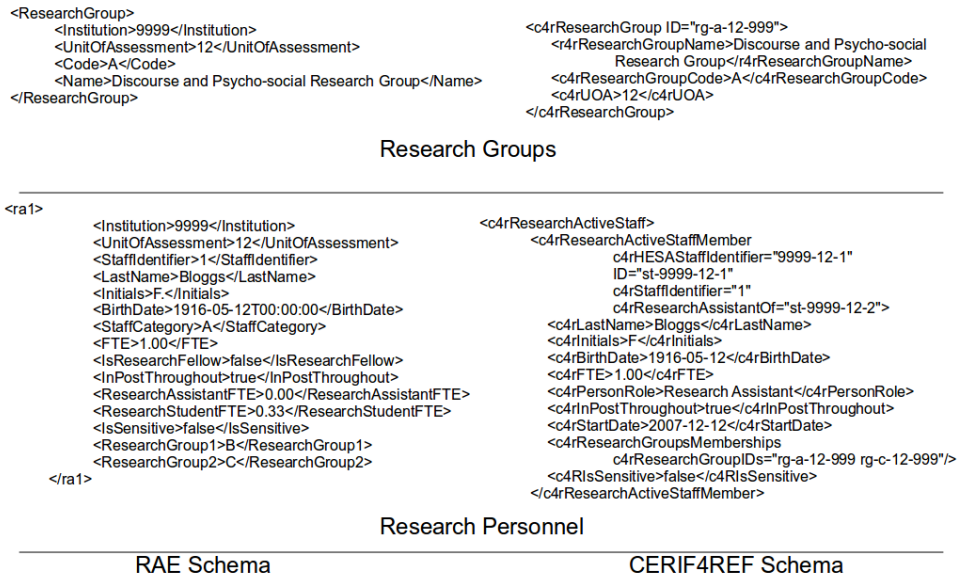


Figure 5. Research Groups and Research Personnel in RAE and CERIF4REF.

Although both encode similar metadata, CERIF4REF relies heavily on XML IDs and IDREFs to establishes linkages between components (for instance, between a research assistant and a supervisor (by use of the **c4rResearchAssistantOf** attribute in this example)); RAE by contrast establishes linkages by element content (for instance the **ResearchGroup1** and **ResearchGroup2** elements). Using IDs in this way allows linkages to be validated with standard XML parsing rather than requiring the use of XCSL or Schematron which the latter approach necessitates. In addition, the contents of many elements in CERIF4REF, for instance **c4rPersonRole** are constrained by closed enumerated lists, and some numerical elements, such as **ResearchAssistantFTE** are absent altogether as this data is calculated directly by aggregating information from elsewhere in the CERIF4REF file (in this case, the number of **c4rResearchAssistantOf** attributes which point to the ID of any given researcher).

The encoding of research outputs also differs between the two schemas:-

Figure 6. Research Outputs in RAE and CERIF4REF.

As is the case for the research group and personnel metadata, the main difference is the increased precision allowed by the more constrained CERIF4REF schema. As before, extensive use is made of XML IDs and IDREFs (here, for instance, linking co-authors and research outputs to their authors and research groups). In addition, concepts which are merged confusingly in the RAE schema can be disentangled so that their components can be encoded more clearly. The RAE schema, for instance, uses the broad element **ShortTitle** to encode journal names when research outputs take the form of journal articles but the

same element is also used to encompass such diverse concepts as volume numbers, patent registration numbers, places of performance or venues for art installations depending on the form of the research output: this conflation of concepts, and their semantic dependence on the value of other elements in the schema (specifically the research output type) introduce a degree of complexity into the encoding process which is likely to be error-prone. The CERIF4REF schema allows for separate elements to be used for each output type, named in a comprehensible manner (for instance **MonographTitle** in the example above) and so greatly reduces the possibility of such errors.

The CERIF4REF schema therefore offers the potential for a simpler and less error-prone medium for encoding the information required by the research assessment exercise for which the earlier RAE schema was devised. To test its practicability, a series of six cases studies were undertaken in which attempts were made to map data held in higher education institutions' digital repositories and human resources, financial and student record systems to CERIF4REF. These studies all produced positive results, indicating that any problems of mapping and exporting to CERIF4REF result from discongruities between the original data requirements of RAE and the ways in which institutions hold and structure data on their systems: no problems were noted that arose through the strategic approach of a constrained, closely internally-linked schema (Gartner 2010, 14).

The second key component to this strategy is the XSLT file which is used to effect the necessary transformations between CERIF4REF and CERIF (and also, should it be required, to RAE). This file is designed to duplicate the mapping function of architectural processing in SGML and also to add a layer of semantics to the resulting XML outputs which was not possible using this earlier methodology. Using certain new features of XSLT 2.0, particularly the **xsl:result-document** element which allows output to be redirected to multiple named files in a single XSLT process, it proved possible to write a single stylesheet to generate all 19 CERIF files which the RAE to CERIF mapping exercise had identified as necessary to encode the range of metadata needed for the research assessment exercise.

In a few cases conversion to CERIF is a simple matter of the extraction and relocation of a given component from the CERIF4REF file to its CERIF equivalent: in this diagram, for instance, the personal identifier **cfPersID** in the CERIF file is a simple translation of the CERIF4REF **c4rHESAStaffIdentifier** attribute of **c4rResearchActiveStaffMember**:-

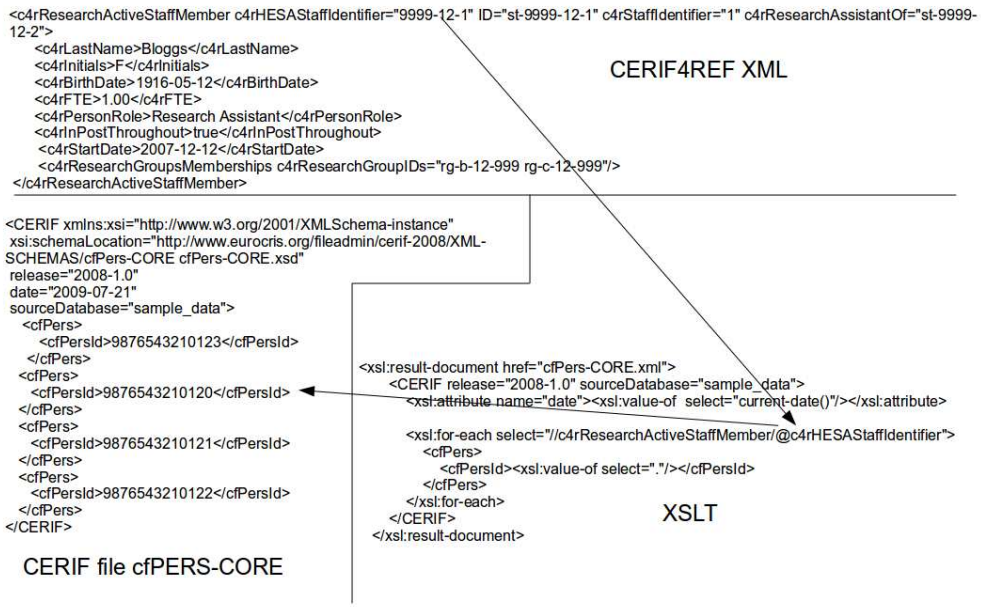


Figure 7. Deriving cfPersId in cfPERS-CORE from CERIF4REF using XSLT.

Producing the linking files which make up the bulk of any CERIF application requires the application of much more complicated XSLT transformations. Figure 8 below shows some of these required to link an individual to a research group:-

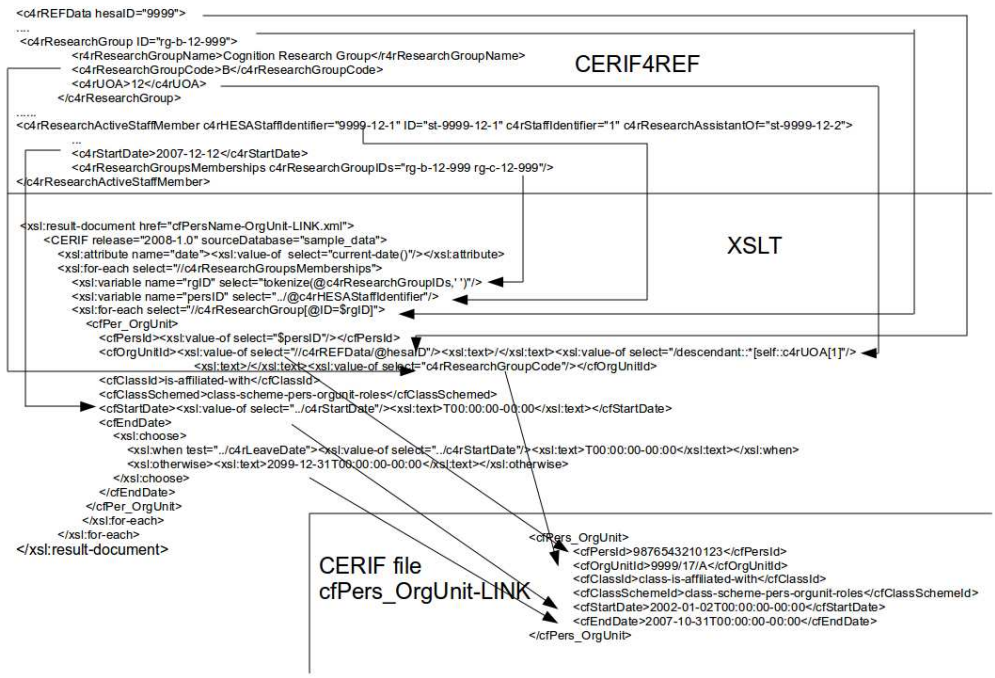


Figure 8. Deriving cfPERS-OrgUnit-LINK from CERIF4REF using XSLT.

The components of the CERIF file here are spread throughout the CERIF4REF file, and some (such as **cfOrgUnitID**) are constructed during processing: the XSLT is therefore of necessity more complex. It is, however, relatively straightforward to write as the CERIF4REF schema was conceived with the mapping to CERIF as its overriding architectural rationale, so that the stylesheet in effect merely translates this mapping to XSLT. It should also be noted that the XSLT file hard-codes the semantic components used by CERIF (in this example **cfClassId** and **cfClassSchemeId**), although these could equally well be supplied by a formal ontology (expressed for instance, in OWL) which is read and processed during the transformations.

Using this approach, CERIF becomes a more viable option for institutions preparing for the forthcoming REF exercise: information held on multiple systems has been shown to map readily to the intermediary schema, and from there, using relatively straightforward stylesheets, it is possible to produce the complex web of CERIF files necessary to express this data and its interrelationships. Reversing the process also proved possible, allowing a stylesheet to be written to extract metadata from multiple CERIF files into the intermediary schema for further editing and re-export to CERIF.

6. Further potential applications

In many metadata environments, particularly in that of the digital library, the problems of complex and highly flexible generic schemas are as acute as they are in that of CERIF. A tension arises particularly between flexibility and interoperability: the more potential approaches to encoding are offered by a standard, the more problematic is the transfer of metadata to other systems and its interpretation and processing by them.

In the digital library arena, this problem has particularly been noted by implementers of the METS (Metadata Encoding and Transmission) (Library of Congress 2011) standard. A widely-read report by the UK's Joint Information Standards Committee (JISC) by the current author (Gartner 2008) proposes METS as the basis of an integrated metadata strategy for digital libraries, but recognises a number of difficulties which arise particularly because of its great flexibility (Gartner 2008, 13-14). Using PREMIS for administrative metadata within METS, for instance, requires the consistent application of best-practice guidelines to resolve such issues as redundancies between the two standards or clashes between the METS's structural metadata functions and those expressed in PREMIS relationships (Gartner 2008, 13).

The usual practice to counteract these problems when implementing METS is to publish a METS Profile in which the usage of METS for a given application is documented in a standardised way (Gartner 2008, 14). Such Profiles, however, are merely human-readable documentation of a METS application, and are not machine-actionable as a mechanism for allowing the ready exchange of METS metadata. The use of intermediary schemas and associated XSLT transformations as documented here offers a possible way of obviating some of these problems: a METS application, including the extension schemas employed and any controlled vocabularies, could potentially be incorporated into such a schema and used to provide a constrained environment within which metadata can be encoded and from which METS files could be generated.

Such a schema would incorporate any required best-practice guidelines (such as those already published for using PREMIS with METS (Library of Congress 2008)), and ensure full conformance with the application's published METS Profile. Writing XSLT files to translate from a METS instance to the intermediary schema, and then from one intermediary schema to one conforming to another METS Profile, may make it possible to translate between METS profiles in a relatively automated manner. Such a scenario would in itself require some work on standardising approaches to the design of such schemas, but the methodology as it is is certainly capable of sustaining such a function.

7. Conclusions

Despite its great power as an encoding mechanism for the complex metadata needs of research environments, CERIF remains relatively underused in the area of research information management. Its flexibility and fragmented architecture in particular can produce significant problems for implementors and reduce its interoperability unless such key components as its semantic infrastructure are standardised between institutions. These problems were experienced more than a decade ago by implementors of such standards as the TEI and were solved by some by using the architectural mapping features of SGML. Without this facility in XML, the solution advocated here can replicate its best features but also add more powerful, non-syntactic features, such as semantic control.

The strategy has been tested thoroughly in several live research information management environments and found to be generally workable: the only problems experienced have proved to be those inherent in the metadata scheme on which the mapping to CERIF was based. The results have proved it to form a good compromise which allows the use of a key standard (with the consequent benefits of wider interoperability) in conjunction with a constrained, project-specific and more easily implemented element set. The successful application of this methodology suggests that it may be beneficial in the wider area of digital library metadata in general, where several key schemas are more easily implemented when constrained it this way.

8. Acknowledgements

The author acknowledges with thanks the assistance of Brigitte Jörg (Deutsche Forschungszentrum für Künstliche Intelligenz) who provided figures 1, 2 and 3, and the Joint Information Systems Committee (JISC) who financed the R4R project.

9. References

- Centre for e-Research (2011) R4R: Readiness for REF. Available at: http://r4r.cerch.kcl.ac.uk/ [Accessed February 21, 2011].
- European Organisation for International Research Information (2010a) CERIF 2008: 1.2 Semantics. Available at: http://www.eurocris.org/Uploads/Web%20pages/CERIF2008/Release_1.2/CERIF2008_1.2_Semantics.pdf [Accessed February 15, 2011].
- European Organisation for International Research Information (2010b) CERIF Introduction. Available at: http://www.eurocris.org/Index.php?page=CERIFintroduction&t=1 [Accessed February 8, 2011].
- European Organisation for International Research Information (2010c) EuroCRIS Web Site. Available at: http://www.eurocris.org/Index.php?page=CERIFreleases&t=1 [Accessed February 4, 2011].
- Gartner, R. (2008) Metadata for digital libraries: state of the art and future directions, JISC. Available at: http://www.jisc.ac.uk/media/documents/techwatch/tsw_0801pdf.pdf [Accessed January 28, 2010].
- Gartner, R. (2010) Readiness for REF (R4R): CERIF4REF mapping case studies. Available at: http://r4r.cerch.kcl.ac.uk/wp-uploads/2010/11/synthesis1.pdf [Accessed March 1, 2011].
- Gartner, R. & Grace, S. (2010) Modelling national research assessments in CERIF. In *Connecting science with society: the role of research information in a knowledge-based society: 10th International Conference on Current Research Information Systems*. pp. 97-105.
- Higher Education Funding Council for England (2010) HEFCE : Research : Research Excellence Framework (REF). Available at: http://www.hefce.ac.uk/research/ref/ [Accessed February 21, 2011].
- Higher Education Funding Council for England (2008) RAE 2008: Research Assessment Exercise. Available at: http://www.rae.ac.uk/ [Accessed February 21, 2011].
- International Standard Organisation (2006) International Standard: ISO/IEC 19757-3: Information technology: Document: Schema Definition Languages (DSDL) Part 3: Rule-based validation: Schematron. Available at: http://standards.iso.org/ittf/PubliclyAvailableStandards/c040833_ISO_IEC_19757-3_2006(E).zip [Accessed February 15, 2011].
- Jacinto, M.H. et al. (2002) Constraint specification languages : comparing XCSL, Schematron and XML-Schemas. In *XML EUROPE 2002: Proceedings*. XML EUROPE. Barcelona: IDEAlliance.
- Jacinto, M.H. et al. (2004) XCSL: XML Constraint Specification Language. *CLEI Electronic Journal*, 6(1: Paper 1), pp.1-29.
- Jeffery, K.G., Lopatenko, A. & Asserson, A. (2002) Comparative study of metadata for scientific information: the place of CERIF in CRISs and scientific repositories. In *Gaining Insight from Research Information*. Sixth International Conference on Current Research Information Systems. University of Kassel: Kassel University Press, pp. 77-87.
- Jörg, B. et al. eds. (2010) CERIF 2008 - 1.2 Full Data Model (FDM): introduction and specification. Available at: http://www.eurocris.org/Uploads/Web%20pages/CERIF2008/Release_1.2/CERIF2008_1.2_FDM.pdf [Accessed March 16, 2011].
- Library of Congress (2008) Guidelines for using PREMIS with METS for exchange. Available at: http://www.loc.gov/standards/premis/guidelines-premismets.pdf [Accessed March 16, 2011].
- Library of Congress (2011) Metadata Encoding and Transmission Standard (METS) Official Web Site. Available at: http://www.loc.gov/standards/mets/ [Accessed March 4, 2011].
- Simons, G.F. (1999) Using architectural forms to map TEI data into an object-oriented database. *Computers and the Humanities*, 33, pp.85-101.
- Simons, G.F. (1998) Using architectural processing to derive small, problem-specific XML applications from large, widely-used SGML applications. In *Summer Institute of Linguistics Electronic Working Papers*. Summer Institute of Linguistics. Available at: http://www.silinternational.org/silewp/1998/006/SILEWP1998-006.html [Accessed February 1, 2011].
- Welty, C. & Ide, N. (1999) Using the right tools: enhancing retrieval from marked-up documents. *Computers and the Humanities*, 33(1-2), pp.59-84.

Article 2

METS as an intermediary schema for a digital library of complex scientific multimedia.

Information Technology and Libraries. 31 (3), 24–35.

METS as an Intermediary Schema for a Digital Library of Complex Scientific Multimedia

Richard Gartner

ABSTRACT

The use of the Metadata Encoding and Transmission Standard (METS) schema as a mechanism for delivering a digital library of complex scientific multimedia is examined as an alternative to the Fedora Content Model (FCM). Using METS as an "intermediary" schema, where it functions as a template that is populated with content metadata on the fly using Extensible Stylesheet Language Transformations (XSLT), it is possible to replicate the flexibility of structure and granularity of FCM while avoiding its complexity and often substantial demands on developers.

METS as an Intermediary Schema for a Digital Library of Complex Scientific Multimedia

Of the many possible approaches to structuring complex data for delivery via the web, two divergent philosophies appear to predominate. One, exemplified by such standards as the Metadata Encoding and Transmission Standard (METS)[1] or the Digital Item Declaration Language (DIDL),[2] relies on the structured packaging of the multiple components of a complex object within "top-down" hierarchies. The second, of which the Fedora Content Model (FCM) is perhaps a prime example,[3] takes the opposite approach of disaggregating structural units into atomistic objects, which can then be recombined according to the requirements of a given application.[4] Neither is absolute in its approach—METS, for instance, allows cross-hierarchy linkages, and many FCM models are designed hierarchically—but the distinction is clear.

Many advantages are validly claimed for the FCM approach to structuring digital data objects. Individual components, not constrained to hierarchies, may be readily reused in multiple representations with great flexibility.[5] Complex interobject relationships may be encoded using semantic linkages,[6] a potentially much richer approach to expressing these than the structural relationships of XML can allow. Multiple levels of granularity, from that of the collection as a whole down to its lowest-level components, can readily be modelled, allowing interobject relationships to be encoded as easily as intercomponent ones.[7]

Such models, particularly the RDF-based Fedora content model, are very powerful and flexible, but can often lead to complexity and consequently considerable demands on system development before they can be implemented. In addition, despite the theoretical interoperability offered by RDF, in practice the exchange and reuse of content models has proved somewhat limited because considerable work is usually required to re-create and validate a content model created elsewhere.[8]

This article examines whether it is possible to replicate the advantages of this approach to structuring data within the constraints of the more rigid METS standard. The data used for this analysis is a set of digital objects that result from biological nanoimaging experiments, the interrelationships of which present complex problems when they are delivered online. The

Richard Gartner (richard.gartner@kcl.ac.uk) is a Lecturer in Library and Information Science, King's College, London.

method used is an unconventional use of a METS template as an intermediary schema;[9] this allows something of the flexibility of the FCM approach while retaining the relative simplicity of the more structured METS model.

A Nanoimaging Digital Library and its Metadata Requirements

The collection analysed for this study derives from biological nanoimaging experiments undertaken at the Randall Division of Cell and Molecular Biophysics at King's College London. Biological nanoimaging is a relatively new field of research that aims to unravel the biological processes at work when molecules interact in living cells; this is done by using optical techniques that can resolve images down to the molecular level. It has particular value in the study of how diseases progress and has great potential to help predict the effects of drugs on the physiology of human organs.

As part of the Biophysical Repositories in the Lab (BRIL) project at King's College London,[10] a digital library is being produced to meet the needs of practitioners of live cell protein studies. Although the material being made available here is highly specialised, and the user base is restricted to a specialist cohort of biologists, the challenges of this library are similar to those of any collection of digital objects: in particular, the metadata strategy employed must be able to handle the delivery of complex, multifile objects as efficiently as, for example, a library of digitized books has to manage the multiplicity of image files that make up a single digital volume.

The digital library itself is hosted on the widely used Fedora repository platform; as a result, it is employing FCM as the basis of its data modelling. The purpose of this analysis is to ascertain whether METS can also be used for the complex models required by this data and to compare its potential viability as an architecture for this type of application with FCM.

A particular challenge of this collection is that the raw images from which it is constituted require combining and processing before they are delivered to the user. A further challenge is that the library encompasses images from a variety of experiments, all of which combine these files in different ways and employ different software for processing them. Some measure of the complexity of these requirements can be gathered from figure 1 below, which illustrates the processes involved in delivering the digital objects for two types of experiments.

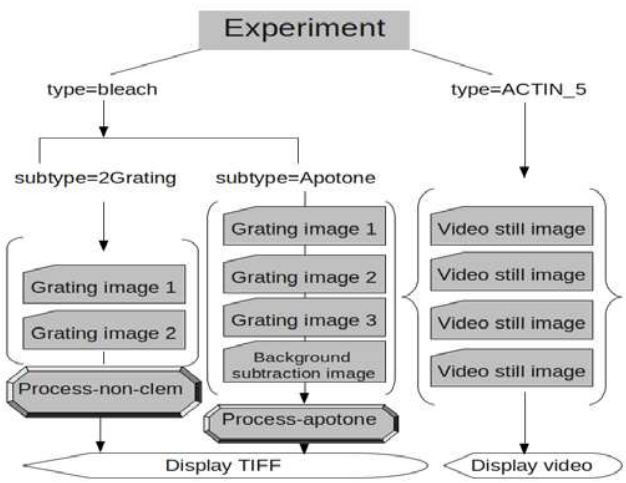


Figure 1. Architecture for Two Experiment Types

The images created by two experiments, *bleach* and *ACTIN_5*, are shown here: it will be seen that the *bleach* experiment is divided into two subtypes (here called *2Grating* and *Apotone*). Each type or subtype of experiment has its own requirements for combining the images it produces before they are displayed.

For the subtype *2Grating***,** for instance, two images, each generated using a different camera grating, are processed in parallel (indicated by the brackets); these are then combined using the software package *process-non-clem* (shown by the hexagonal symbol) to produce a display image in TIFF format. The subtype *Apotone* requires three grating images and a further image with background information to be processed in parallel by the software *process-apotone*; in this case, the background image provides data to be subtracted from the combined three grating images to produce the final TIFF for display. *ACTIN_5* experiments are entirely different: they produce still images that need to be processed sequentially (shown by the braces) to produce a video.

Encoding the BRIL Architecture in METS

This architecture, although complex, is readily handled within METS in a manner analogous to that of more conventional collections. As in any METS file, the structure of the experiments, including their subexperiments, is encoded using nested division (* *) elements in the structural map (example 1a).

```
<div TYPE= “experiment-types”>
<div TYPE= “experiment-bleach” ID= “experiment-type0001”>
<div TYPE= “experiment-bleach-2Grating” ID= “experiment-type0001-subtype001”>
 [subsidiary <div>s containing image information]
</div>
<div TYPE= “experiment-bleach-apotone” ID= “experiment-type0001-subtype002”>
 [subsidiary <div>s containing image information]
</div>
</div>
<div TYPE= “experiment-actin_5” ID= “experiment-type0002”>
 [subsidiary <div>s containing image information]
</div>
</div>
```

Example 1a. Sample Experiment-Level Structural Map

Within these containing divisions, subsidiary * * elements are used to map the combination of images necessary to deliver the content for each type. METS allows the specification of the parallel or sequential structuring of files using its *<par>* and *<seq>* elements respectively. The parallel processing of the *Apotone* subtype, for instance, could be encoded as shown in example 1b.

```
<div TYPE= “experiment-bleach-apotone-images” ID= “apotone-000”>
<fptr>
 <par>
 <area FILEID= “apotone-grating1–000”/>
 <area FILEID= “apotone-grating2–000”/>
 <area FILEID= “apotone-grating3–000”/>
 </par>
</fptr>
</div>
```

Example 1b. Sample Parallel Structure for Raw Image Files to be Combined Using a Process Specified in Associated Metadata (Behavior Section)

Each division of the structural map of this type may in its turn be attached to a specific software item in the METS behavior section to designate the application through which it should be processed: the tri-partite set of images in example 1b, for instance, would be linked to the *process-apotone* software using the code in example 1c.

```
<behavior GROUPID= “process-apotone” STRUCTID= “apotone-000”>
 <mechanism LOCTYPE= “URL” ns2:href= “software/process-apotone”/>
</behavior>
```

Example 1c. Sample METS Behavior Mechanism for a Specification of Image Processing

This approach is straightforward, and METS is capable of encoding all of the requirements of this data model, although at the cost of large file sizes and a degree of inflexibility. This may be no problem when the principle rationale behind the creation of this metadata is preservation: linking all of the project metadata in a coherent, albeit monolithic, structure of this kind benefits especially its usage as an Open Archival Information System (OAIS) Archival Information Package (AIP), one of the key functions for which METS was designed. Problems are likely to arise, however, when this approach is scaled up in a delivery system to include the potentially millions of data objects that this project may produce.

The large size of the METS files that this approach necessitates makes their on-the-fly processing for delivery much slower than a system that uses aggregations of the smaller files required by the FCM model and so processes only metadata at the granularity necessary for the delivery of each object. Such flexibility is much harder to achieve within METS, although mechanisms that currently exist for aggregating diverse objects within METS may seem to offer some degree of solution to this problem.

Complex Relationships under METS

Underlying the METS structural map is an assumed ontology of digital objects that encodes a long-established view of text as an ordered hierarchy of content objects;[11] this model accounts for the

map's use of hierarchical nesting and the ordinality of the object's components. The rigidity of this model is alleviated to some extent by the facility within METS to encode structural links that cut across these hierarchies. These links, which join nodes at any level of the structural map, are particularly useful for encoding hyperlinks within webpages,[12] and so are often used for archiving websites.

Various attempts have been made to extend the functionality of the structural map and structural links sections to allow more sophisticated aggregations and combinations of components beyond the boundaries of a single digital object, in a manner analogous to the flexible granularity of FCM. METS itself offers the possibility of aggregating other METS files through its *<mptr>* (METS pointer) element: this element, always a direct child of a * * element in the structural map, references a METS file that contains metadata on the digital content represented by this * *. For example, two complex digital objects could be represented at a higher collection level, as shown in example 2.

```
<div>
<div ID= "bleach2grating-1">
<mptr LOCTYPE= "URL" xlink:href= "twogratingbleach6-grating1–01.xml"/>
</div>
<div ID= "bleach2grating-2">
<mptr LOCTYPE= "URL" xlink:href= "twogratingbleach6-grating1–02.xml"/>
</div>
</div>
```

Example 2. Use of METS *<mptr>* Element

This feature has found some use in such projects as the ECHO DEPository, which uses it to register digital objects at various stages of their ingest into, and dissemination from, a repository;[13] it is also recommended by the PARADIGM project as a method for archiving born-digital content, such as emails.[14]

Nonetheless, its usage remains fairly limited; of all the METS Profiles registered on the central METS repository, for instance, ECHO DEP at the time of writing remains the only project on the Library of Congress's repository of METS Profiles to employ this feature. [15] An important reason for its limited take-up is that its potential for more sophisticated uses than merely populating a division of the structural map is severely limited by its place in the METS schema. The *<mptr>* element can only be used as a direct child of its parent * *: it cannot, for instance, be located in *<par>* or *<seq>* elements to indicate that the objects referenced in its subordinate METS files should be processed in parallel or in sequence (as is required by the different experiment types in figure 1), nor may the contents of these files be processed by the sophisticated partitioning features of the *<area>* element, which allows subsidiary parts of a * * to be addressed directly.

A more sophisticated approach to combining digital object components is to employ Open Archives Initiative Object Reuse and Exchange (OAI-ORE) aggregations,[16] which express more complex relationships at greater levels of granularity than the *<mptr>* method allows.

McDonough's examination of the possibility of aligning the two standards concludes that it is indeed possible, although at the cost of eliminating the METS behavior section and removing much of the flexibility of METS's structural links, both side effects of OAI-ORE's requirement that resource maps must form a connected RDF graph.[17] In addition, converting between METS and OAI-ORE may not be lossless, depending on the design of the METS document.[18]

Neither approach therefore seems ideal for an application of this type, the former because of the limited ways in which the *<mptr>* element can be deployed outside the *<area>* element and its subsidiaries, the latter because of its removal of the functionality of the behavior section, which is essential for the delivery of material such as this.

METS as an Intermediary Schema

An alternative approach adopted here uses the technique of employing METS files as intermediary schemas to act as templates from which METS-encoded packages for delivery can be generated. Intermediary XML schemas are intermediary in the sense that they are designed not to act as final metadata containers for archiving or delivery, but as mediating encoding mechanisms from which data or metadata in these final forms can be generated by XSLT transformations: one example is CERIF4REF, a heavily constrained XML schema used to encode research management information from which metadata in the complex Common European Research Information Format (CERIF) data model can be generated.[19]

The CERIF4REF schema attempts to emulate the architectural processing features of SGML,[20] which are absent from XML; these allowed simpler Document Type Definitions (DTDs) to be compiled for specific applications, which could then be mapped to established, more complex, SGML models. Instead of architectural processing, CERIF4REF uses XSLT to carry out this processing, so allowing the combination of a simpler scheme tailored to the requirements of an application to be combined with the benefits of a more interoperable but highly complex model that is difficult to implement in its standard form.

Instead of using this technique for constraining encoding to a simpler model and generating more complex data structures from this, the intermediary schema technique may be used to define templates, similar to a content model, from which the final METS files to be delivered can be constructed. As is the case with CERIF4REF, XSLT is used for these transformations, and the XSLT files form an integral part of the application. In this way, a series of templates, beginning with highest-level abstractions, are used to generate their more concrete subsidiaries, until a final version used for dissemination is generated. The core of this application is a METS file, which acts as a template for the data delivery requirements for each type of experiment. Figure 2 demonstrates the components necessary for defining these for the *2Grating* experiment subtype detailed previously in figure 1.

Figure 2. Defining an Experiment Subtype in METS

The data model for the delivery of these objects is defined in the *<structMap>* (b): as can be seen here, a series of nested * * elements is used to define the relationship of experiment subtypes to types, and then to define, at the lowest level of this structure, the model for delivering the objects themselves. In this example, two files are to be processed in parallel; these are defined by *<area>* elements within the *<par>* (parallel) element. In a standard METS file, the *FILEID* attribute of *<area>* would reference a *<file>* element within the METS file section (a): in this case, however, they reference empty file group *(<fileGrp>)* elements, which are populated with *<file>* elements when this template undergoes its XSLT transformation.

The final component of this template is the METS behavior section (c), in which the applications required to process the digital objects are defined. Two behavior sections are shown in this example: the first is used to invoke the XSLT transformation by which this METS template file is to be processed, the second to define the software necessary to co-process the two images files for delivery. Both indicate the divisions of the structural map whose components they process by their *STRUCTID* attributes: the first references the map as a whole because it applies to recursively to the METS file itself, the second references the experiment for which it is needed.

When delivering a digital object, it is then necessary to process this template METS file to generate the final version used to encode its metadata in full. The XSLT used to do this co-processes the template and a separate METS file defined for each object containing all of its relevant metadata:

this latter file is used to populate the empty sections of the template, in particular the file section. Figure 3 provides an illustration of the XSLT fragment which carries out this function.

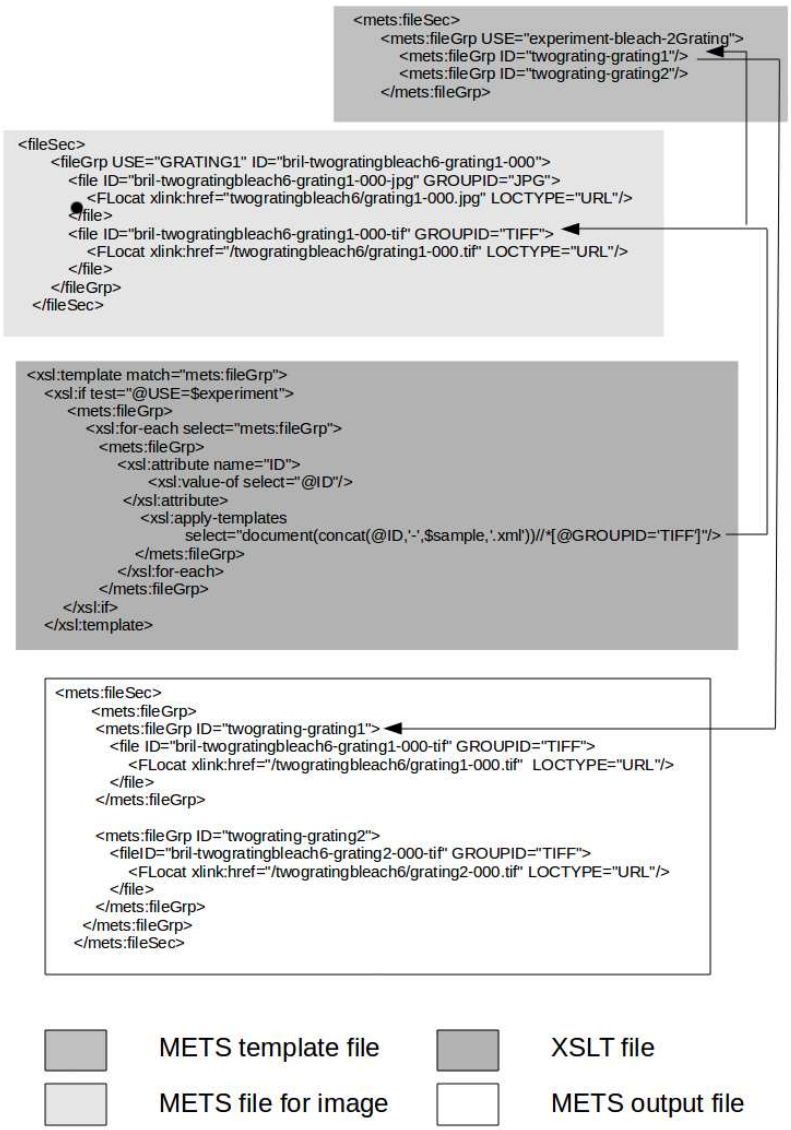


Figure 3.

The XSLT transformation file is evoked with the *sample* parameter, which contains the number of the sample to be rendered: this is used to generate the filename for the *document* function, which selects the relevant METS file containing metadata for the image itself. The *<file>* element within this file, which corresponds to the required image, is then integrated into the relevant *<fileGrp>* element in the template file, populating it with its subcomponents, including the *<FLocat>* element, which contains the location of the file itself.

In the case of the *ACTIN_5* experiment, which generates a video file from a sequence of still images, the processes involved are slightly more complicated. Because the number of still images to be processed will vary for each sample, it is not possible to specify the template for the delivery

version of the sequence explicitly within a *<fileGrp>* element as is done for the other experiments. Instead, it is necessary to define a further METS file (the "sequence file") in which the sequence for a given sample is defined. In this case, the architecture is shown in figure 4.

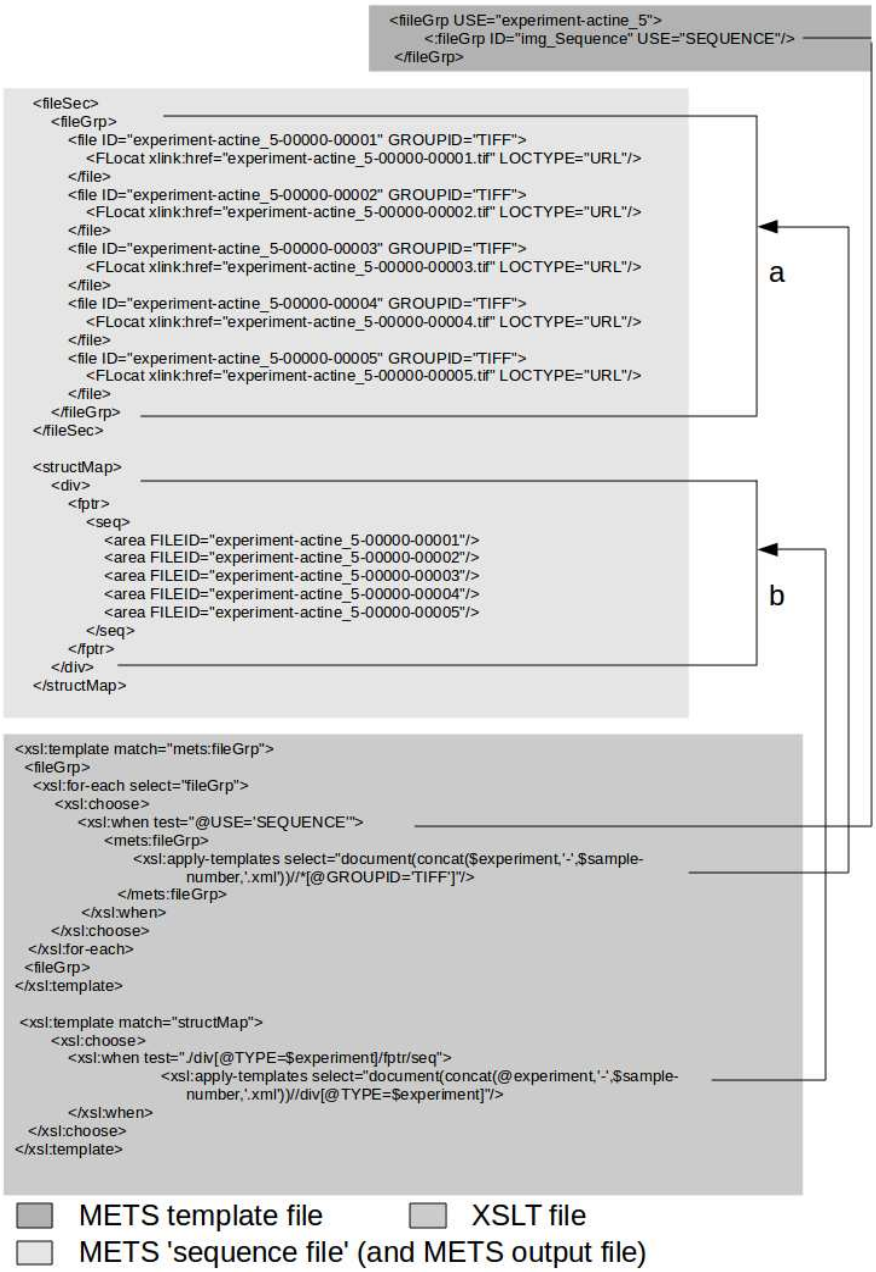


Figure 4. Populating Sequentially Processed File Section with XSLT

In this case the *<fileGrp>* element in the METS template file acts as a placeholder only and does not encode even the skeletal information for the parallel-processed TIFF files in figure 3. Similarly, the structural map * * for this experiment indicates only that this section is a sequence but does not enumerate the files themselves even in template form. Both of these sections are populated when the file is processed by the XSLT transformation to import metadata from the METS "sequence file,"

in which the file inventory (a) and sequential structure (b) for a given sample are listed. The XSLT file populates the file section and structural map directly from this file, replacing the skeletal sections in the template with their counterparts from the sequence file.

Through this relatively simple XSLT transformation, the final delivery version of the METS file is readily generated for either content model. This file can itself then be delivered on the fly (for instance, as a Fedora disseminator); this is done by using a further XSLT file to process the complex digital object components using the mechanism associated with each experiment in the METS behavior section. Given the relatively small size of all of the files involved, this processing can be done more quickly than would be possibly using a fully aggregated METS approach. In the laboratory environment in particular, where the fast rendering and delivery of these images is needed so as not to impede workflows, this has major advantages.

Although the project aimed to examine specifically the use of Fedora for the delivery of this complex material, and so employed FCM as the basis of its metadata strategy, the technique examined in this article proved itself a viable alternative that made much fewer demands on developer time. The small number of XSLT stylesheets required to render and deliver the METS files were written within a few hours: the development time to program the delivery of the RDF-based metadata that formed the FCM required several weeks. Processing XML using XSLT disseminators in Fedora is very fast, and so using this method instead of processing RDF introduces no discernible delays in object delivery.

CONCLUSIONS

This approach to delivering complex content appears to offer the benefits of the alternative approaches outlined above in a simpler manner than either currently allows. It offers much greater flexibility than the METS <*mptr*> element, which can only populate a complete structural map division. When compared to the FCM approach, this strategy, which relies solely on relatively simple XSLT transformations for processing the metadata, requires less developer time but offers a similar degree of flexibility of structure and granularity. It also avoids much of the rigidity of the OAI-ORE approach by not requiring the use of connected RDF graphs, and so frees up the behavior section to define the processing mechanisms needed to deliver these objects.

Using the intermediary schema technique in this way does therefore offers a means of combining the advantages of employing well-defined interoperable metadata schemes and the practicalities of delivering digital content in an efficient manner, which makes limited demands on development. As such, it represents a viable alternative to the previous attempts to handle complex aggregations within METS discussed above.

The adoption of integrated library systems (ILS) became prevalent in the 1980s and 1990s as libraries began or continued to automate their processes. These systems enabled library staff to work, in many cases, more efficiently than they had been in the past. However, these systems were also restrictive—especially as the nature of the work began to change—largely in response to the growth of electronic and digital resources for which they were not intended to manage. New library systems—the second (or next) generation—are needed to effectively manage the processes of acquiring, describing, and making available all library resources. This article examines the state of library systems today and describes the features needed in a next-generation ILS. The authors also examine some of the next-generation ILSs currently in development that purport to fill the changing needs of libraries.

REFERENCES

[1] Library of Congress, "Metadata Encoding and Transmission Standard (METS) Official Web Site," 2011 http://www.loc.gov/standards/mets (accessed August 1, 2011).

[2] Organisation Internationale de Normalisation, "ISO/IEC JTC1/SC29/WG11: Coding of Moving Pictures and Audio," 2002, http://mpeg.chiariglione.org/standards/mpeg-21/mpeg-21.htm (accessed August 1, 2011).

[3] Fedora Commons, "The Fedora Content Model Architecture (CMA)," 2007, http://fedora-commons.org/documentation/3.0b1/userdocs/digitalobjects/cmda.html (accessed December 9, 2011).

[4] Carl Lagoze et al., "Fedora: An Architecture for Complex Objects and their Relationships," *International Journal on Digital Libraries* 6, no. 2 (2005): 130.

[5] Ibid., 127.

[6] Ibid., 135.

[7] Ibid.

[8] Rishi Sharma, *Fedora Interoperability Review* (London: Centre for e-Research, 2007), http://wwwcache1.kcl.ac.uk/content/1/c6/04/55/46/fedora-report-v1.pdf.3 (accessed August 1, 2011).

[9] Richard Gartner, "Intermediary Schemas for Complex XML Publications: An Example from Research Information Management," *Journal of Digital Information* 12, no. 3 (2011), https://journals.tdl.org/jodi/article/view/2069 (accessed August 1, 2011).

[10] Centre for e-Research, "BRIL," n.d., http://bril.cerch.kcl.ac.uk (accessed August 1, 2011).

[11] S. J. DeRose et al., "What is Text, Really," *Journal of Computing in Higher Education* 1, no. 2 (1990): 6.

[12] Digital Library Federation, "<METS>: Metadata Encoding And Transmission Standard: Primer And Reference Manual," Digital Library Federation, 2010, www.loc.gov/standards/mets/METSPrimerRevised.pdf, 77 (accessed August 1, 2011).

[13] Bill Ingram, "ECHO Dep METS Profile for Master METS Documents," n.d., http://dli.grainger.uiuc.edu/echodep/METS/DRAFTS/MasterMETSProfile.xml (accessed August 1, 2011).

[14] Susan Thomas, "Using METS for the Preservation and Dissemination of Digital Archives," n.d., www.paradigm.ac.uk/workbook/metadata/mets-altstruct.html (accessed August 1, 2011).

[15] Library of Congress. "METS Profiles: Metadata Encoding and Transmission Standard (METS)

OfficialWeb Site", 2011. http://www.loc.gov/standards/mets/mets-profiles.html (accessed December 6, 2011).

[16] Open Archives Initiative, "Open Archives Initiative Protocol—Object Exchange and Reuse," n.d., www.openarchives.org/ore (accessed December 12, 2011).

[17] Jerome McDonough, "Aligning METS with the OAI-ORE Data =Mmodel," *JCDL '09 Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital libraries* (New York: Association for Computing Machinery, 2009): 328.

[18] Ibid., 329.

[19] Gartner, "Intermediary Schemas."

[20] Gary Simons, "Using Architectural Processing to Derive Small, Problem-Specific XML Applications from Large, Widely-Used SGML Applications," *Summer Institute of Linguistics Electronic Working Papers* (Chicago: Summer Institute of Linguistics, 1998), www.silinternational.org/silewp/1998/006/SILEWP1998-006.html (accessed August 1, 2011).

Article 3

The digital object in context: using CERIF with METS.

Journal of Library Metadata. 12 (1), 39–51.

This is an electronic version of an article published in Journal of Library Metadata, 12(1), pp. 39-51. Journal of Library Metadata is available online at: http://www.tandfonline.com/doi/full/10.1080/19386389.2012.661689

The digital object in context: using CERIF with METS

Running title: The digital object in context

The article examines the potential for using the Common European Research Information Format in conjunction with the Metadata Encoding and Transmission Standard to provide contextual information for a digital research output. Both are key standards within their respective communities (the former in research information management, the latter in digital library metadata), but each employs a different approach to information architecture which renders their combination problematic. The article examines three possible ways to using CERIF and METS in conjunction, and suggests possible changes to approach of the METS standard to resolve some of the problems that arise.

Introduction

A report in the Technology and Standards Watch series published by the UK's Joint Information Systems Committee (Gartner 2008) argues for an integrated metadata strategy for digital library objects based on the METS (Metadata Encoding and Transmission Standard) XML schema (Library of Congress 2011a) . The key advantage of using METS which this report cites is its ability to incorporate external XML schemas (extension schemas) for descriptive or administrative metadata while maintaining a single overall architecture (Gartner 2008, p.15); this, it is argued, provides a degree of integration whilst maintaining the flexibility of using the metadata schemes most relevant to a given application.

Common extension schemas used with METS include MODS (Metadata Object Description Schema) for descriptive (Library of Congress 2010) and PREMIS

(PREservation Metadata: Implementation Strategies) for preservation metadata (Library of Congress 2011b). These have proved themselves highly effective in meeting the needs of the users and administrators of digital libraries, but in the case of digital objects which are themselves research outputs (for instance, journal articles held in an institutional repository) further contextual information on the research environments in which they were created may be particularly valuable in critically evaluating their contribution to the academic record.

A standard for encoding information of this type has been existence for over ten years in the form of the CERIF (Common European Research Information Format) data model (European Organisation for International Research Information 2010). This scheme is now well established, particularly in Europe, as the core standard for interoperable research information management; it is, for instance, the recommended format for exchanging this data in the United Kingdom (Rogers, Huxley & Ferguson 2010, p.23), and will be a submission format for the UK's next national research assessment exercise for higher education (Bolton 2010, p.21).

This article argues that there is a strong case for using CERIF within METS as an adjunct to more standard metadata schemes in order to provide important contextual information for a digital object. However, the complexity of the CERIF standard, which is based on a data model derived from relational database tables, raises questions as to the usability of METS's currently limited facilities for addressing the contents of the metadata it embeds or references. Although an approach to resolving this within the current METS framework of metadata 'buckets' is discussed here, it is arguable that METS needs to be extended to allow more sophisticated approaches to metadata in general.

The research object in context

The need to provide contextual information on a research output is inherent in the concept of the Current Research Information System (CRIS) which holds information on the projects, organisations, people and funding that contribute to the production of the output itself (Jeffery et al. 2002, p.78). Such information is important to a wide range of potential stakeholders in the research environment. Most obviously, funding authorities require this information for allocating resources: the UK's 2008 Research Assessment Exercise (RAE), for instance, on the basis of which much of its research funding to higher education institutions was distributed, required a detailed set of information on such factors as funding sources, studentships, numbers of research students, research groupings and the overall research environment (Higher Education Funding Council for England 2007) in order to allow a valid assessment of the value of the outputs submitted for evaluation.

For the individual researcher, contextual information of this type may also be valuable in assessing the relevance and potential value of their peers' outputs. A researcher may, for instance, want to evaluate a paper in the sciences by finding out what software was used in the experiment it documents or by examining the original raw data on which the paper is based (Jeffery, Lopatenko & Asserson 2002, p.79). A counterpart in the humanities or social sciences may perhaps attempt to deduce the limitations or potential biases in a work by examining the context of the institutional setting in which it was produced (and its possible effects on research agendas) or by examining the funding sources which supported its creation. In current environments where the overall direction of so much research is dictated by the priorities of funding bodies, information of this type can be invaluable as an adjunct to more standard methods for assessing the relevance of a work to the individual's own research.

A metadata environment for the digital research output

Where the output of research is available in digital form, particularly as part of an institutional repository, similar metadata requirements apply as in any digital library environment. In particular, a packaging standard is necessary to structure the often complex diversity of descriptive, administrative and structural metadata necessary to ingest, maintain and deliver the object. The most obvious choice for this is METS which is already well established as standard of this type for digital library metadata. METS can function as a Submission Information Package (SIP), Archival Information Package (AIP) or Dissemination Information Package (DIP) under the OAIS (Open Archival Information System) (Consultative Committee for Space Data Systems 2009), and so can be used throughout the submission, archiving and delivery chain for digital objects. It has also been used successfully in institutional repositories, and is supported by such key repository systems as ePrints (University of Southampton 2010), Dspace (Massachusetts Institute of Technology 2010) and Fedora (Fedora Project 2006). For all of these reasons, using METS as the underlying architecture for an integrated metadata strategy for digital research objects is a sensible option.

The choice of CERIF for providing the contextual information for a research object is equally compelling. This standard provides the only comprehensive, interoperable mechanism for such metadata, and has now gained sufficient traction as an approved standard within higher education sectors to be regarded as the only viable option for this.

Using METS and CERIF in conjunction therefore appears to be a sensible option for a digital repository or CRIS: the former brings all of the advantages of a coherent packaging standard which is now well established in the digital library community, the latter the ability to contextualise the digital

research output using an interoperable and highly flexible scheme which has established itself as the lingua franca for research information management.

Using CERIF for research information management

CERIF was initially published as a data model realised in a set of relational SQL tables. Since 2006, it has also been made available as a set of 192 XML schemas which duplicate the architecture of the SQL original. The CERIF model in both SQL and XML defines a small set of entities and then provides an extensive set of linking mechanisms to realise their relationships in a working environment.

The core of CERIF is a set of very basic 'base' information entities, which document projects, people and organisational units: for projects, for instance, the information recorded here limited to fields for an internal ID, URI, acronym, start and end dates, title, abstract and keywords. This core set is supplemented by a further small set of 'result' entities which record information on research outputs (publications, patents and products). These again include only basic components: for publications, for instance, only the core metadata required to identify a work (such as identification numbers, date and pagination) is included here.

A further set of 16 entities, termed "second-level" in the CERIF specification, includes an array of subsidiary concepts which may be used to qualify project, person, organisational unit or result entities: these include such components as metrics, events, qualifications, facilities, equipment or expertise. Another set of 49 entities handles the multi-lingual features of a CERIF application, allowing multiple language versions of any textual information to be encoded: for example, multiple language versions of the title of this article would be rendered as follows:-

<cfResPublTitle>

<cfResPublId>1abc</cfResPublId>

<cfTitle cfLangCode="en-UK">The digital object in context: using CERIF with METS</cfTitle>

<cfTitle cfLangCode="de">Das digitale Objekt im Kontext: CERIF mit METS benutzen</cfTitle>

<cfTitle cfLangCode="it">L'oggetto digitale nel contesto: utilizzando CERIF con METS</cfTitle>

</cfResPublTitle>

The core of the CERIF application is a set of 95 linking entities which mirror the relational database linkages of the original data model and allow the base, result, second-level and multi-lingual entities to be joined together. These entities usually contain nothing more than the IDs of the two components being linked, the semantic terms which establish the nature of the link itself and start and end dates for its validity: for instance, linking a researcher to their institutional affiliation may be done as shown below, where the cfClassId element contains the semantic term and the cfClassSchemeId indicates the vocabulary from which it is taken:-

<cfPers_OrgUnit>

<cfPersId>9876543210123</cfPersId>

<cfOrgUnitId>9999/17/A</cfOrgUnitId>

<cfClassId>class-is-affiliated-with</cfClassId>

<cfClassSchemeId>class-scheme-pers-orgunit-roles</cfClassSchemeId>

<cfStartDate>2002-01-02T00:00:00-00:00</cfStartDate>

<cfEndDate>2007-10-31T00:00:00-00:00</cfEndDate>

</cfPers_OrgUnit>

All cfClassIds and cfClassSchemeIds, and their associations, are in their turn defined in "Class" schemas.

A CERIF application therefore requires the use of a substantial number of XML instances, each conforming to its own schema; these are then linked together by a coherent system of IDs and a series of semantic vocabularies.

METS and CERIF: possible approaches to integration

Several possible approaches could feasibly be adopted to allow CERIF to be integrated with the METS framework. The simplest is to use CERIF as an extension schema in a manner analogous to MODS or PREMIS and embed or reference its component XML files from within METS's metadata sections. A further approach may be to embed a simpler and more constrained XML file conforming to an 'intermediary' schema from which the CERIF files could be derived. A third approach, which perhaps extends METS beyond its intended functionality, would be to attempt to serialize the relationships expressed within the CERIF files in the METS structural map.

CERIF as an extension schema

Because CERIF is available in XML form, it is feasible to either embed all of the CERIF files associated with a digital research output within METS <mdWrap> elements or hold them externally and reference them from within <mdRef> elements within the <dmdSec> or <admSec> sections. This is by far the simplest approach and fits cleanly within the METS model for extension schemas.

An initial problem with this approach is a degree of obscurity as to where in METS's architecture the CERIF files belong. Although some, such as those which provide bibliographic information akin to such standards as MODS or MARC, fit neatly into METS's <dmdSec> element for descriptive metadata, the majority of files, particularly the linking components which form the majority of a CERIF application, do not fall so readily into the METS framework. These files may perhaps be rationally considered part of an object's administrative metadata, but few fit neatly into the four sub-components (technical, source, rights or digital provenance) into which METS's section for this type of metadata, <amdSec>, is divided.

The problems of fitting some types of metadata into the METS taxonomy have been noted by a number of implementers (METS Editorial Board 2010, p.11), but at present METS's requirement that metadata (in particular administrative metadata) is subdivided into separate "buckets" prevents any clean fit with CERIF. Although unsatisfactory in many ways, for the purpose of this article CERIF files will be put into METS <dmdSec> elements, an approach analogous to that taken by Habing and Cole (2008) in encoding aggregation RDF metadata within METS.

More difficulties arise from the complexity of the CERIF standard itself. As shown above, a CERIF application may involve a total of 192 separate files linked together by a complex, and highly flexible, system of semantic classes. Navigating this set of linkages in order to extract relevant metadata and present it in a meaningful way requires complicated processing and extensive documentation of the data architecture employed in a given CERIF application. These problems become more acute when the METS framework is intended to facilitate interoperability between systems. The use of METS for digital object interchange has been acknowledged as one of its less successful features and some writers, such as Maslov, claim with some reason that it can only function as a packaging rather than exchange protocol as it lacks the

specificity needed to generate an unambiguous interpretation of its encoded data and metadata (Maslov et al. 2010).

The standard mechanism for facilitating exchange between METS applications is the METS Profile, which documents key features, such as the external schemas and structural requirements, of an application in a standardised way. It are not, however, intended to be machine actionable: all parts are merely human-readable descriptions which must be actioned by those responsible for the design of the receiving system. Consequently, it prone to inconsistencies and idiosyncracies, and lacks the specificity which Maslov et al. point to as an impediment to interoperability.

These problems become particularly acute when an application as complex as CERIF is incorporated into METS. Interpreting the possible 192 files that make up an application requires a much greater degree of precision of documentation than a simpler, unitary schema such as MODS necessitates, and even careful recording of this type within a METS Profile will require an intimidating degree of analysis and development by the administrators of a system ingesting a CERIF-enabled METS file before it can sensibly be actioned.
While this simple use of METS as an extension schema is, therefore, entirely feasible when it is intended to function as a packaging protocol alone, it may prove more problematic when it is intended to transfer objects between systems.

Use of an 'intermediary' schema

An alternative approach to the handling the complexity of CERIF was adopted by the UK's Readiness for REF (R4R) (Centre for e-Research 2011) project in the form of 'intermediary' XML schemas. These are heavily constrained schemas which cut down a complex and flexible model such as CERIF to a simplified form designed specifically for a given application. XSLT transformations are then

used to generate the required CERIF files with consistent semantics and linkages (Gartner 2011). The R4R project created a bespoke schema of this kind, CERIF4REF, designed specifically to produce CERIF-compliant submissions to the UK's periodic research assessment exercise.

Adopting this approach would entail embedding or referencing an XML instance conforming to such an 'intermediary' schema within a METS <dmdSec> rather than the CERIF files themselves. A <behavior> element within the METS behavior section would then contain a <mechanism> element which would reference the associated XSLT needed to generate the raw CERIF files. The STRUCTID attribute of the <behavior> element would reference a element in the structural map which in turn references an entry in the <fileSec> that points to the <dmdSec> containing the intermediary XML instance; this requires a recursive referencing of the <dmdSec> from the href attribute of the <FLocat> element, an approach similar to that previously advocated by Habing & Cole (2008) as a method for structuring OAI-ORE aggregation data within the METS framework.

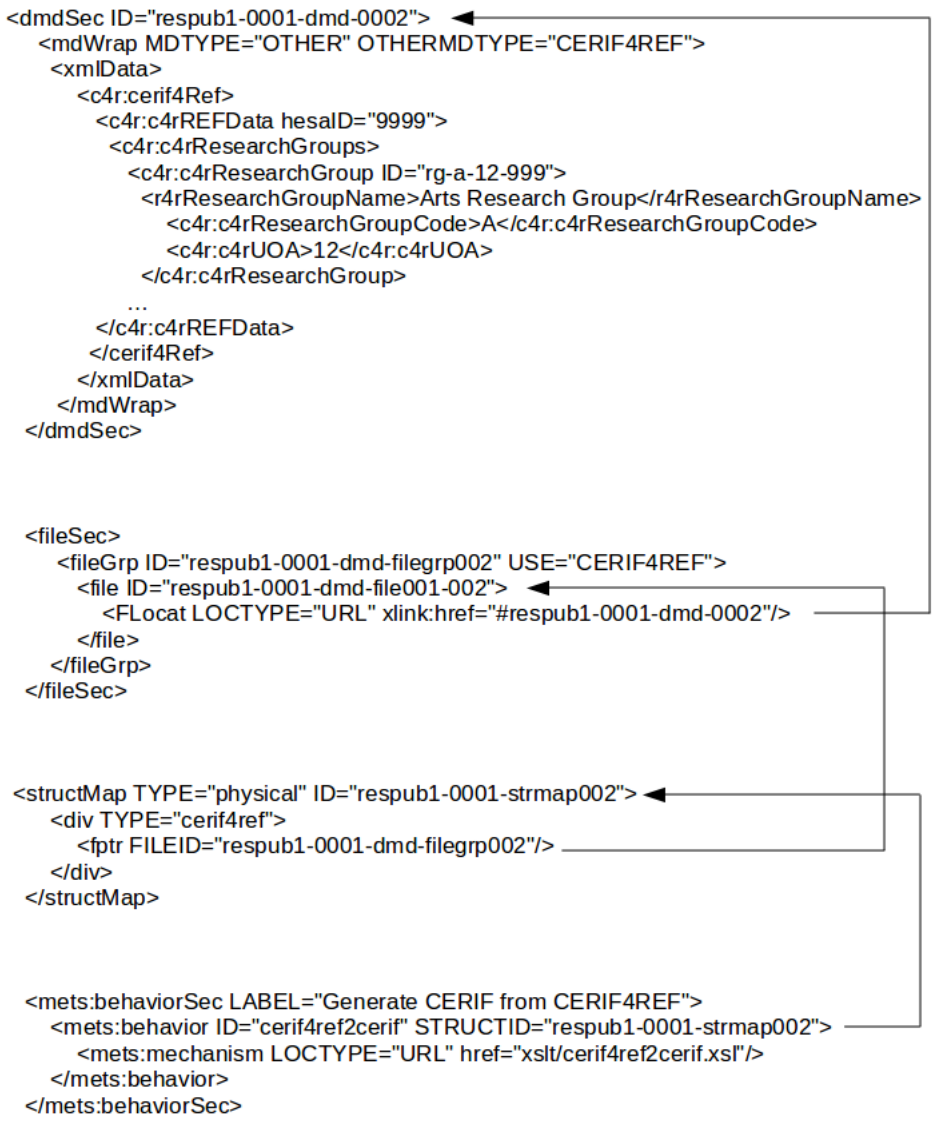


Figure 1 illustrates this chain of references.

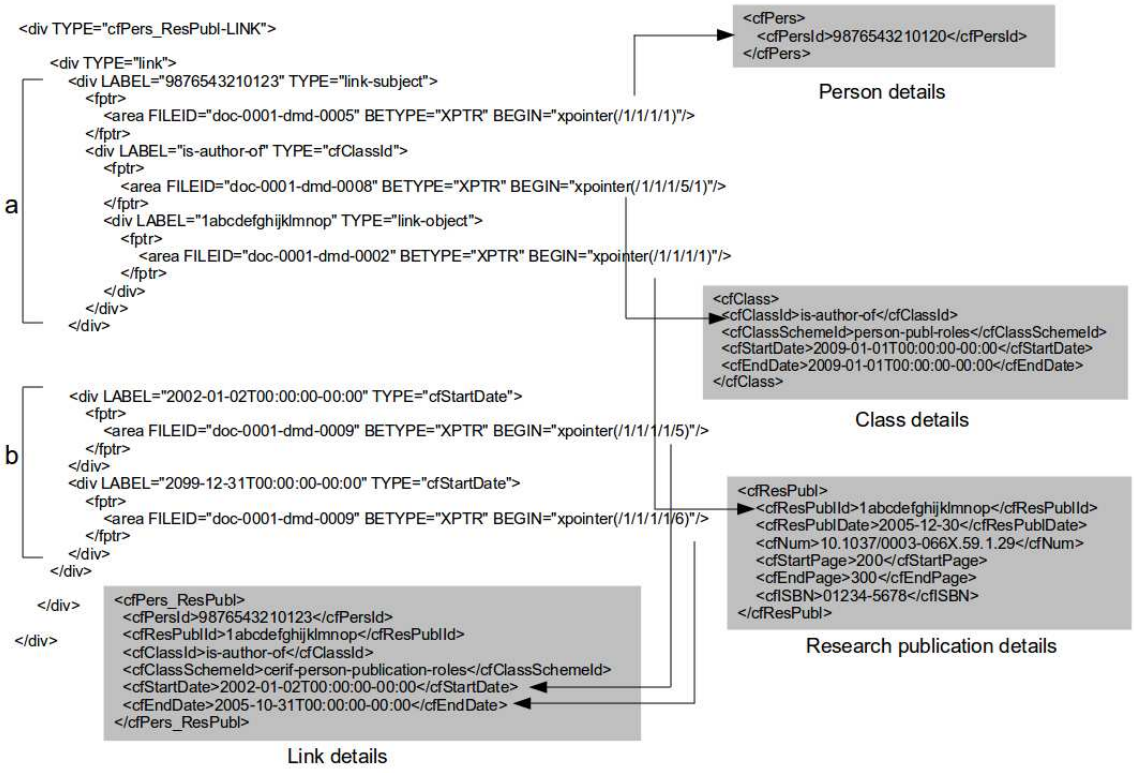
This approach should, if the intermediary schema is well designed, substantially reduce the complexity of the CERIF application and the consequent difficulties in designing suitable mechanisms for rendering its constituent components for delivery. The exchange of METS files containing contextual metadata encoded in this way should also be considerably easier, and require less documentation and interpretation by the recipient system. Such an approach does, however, involve the use of a bespoke, non-standard, schema of limited use beyond the project for which it is designed. Although such schemas would almost certainly be simpler than CERIF itself, a proliferation of unstandardised schemas of this kind would nullify the

benefits of employing a standard for encoding and exchanging research information, and probably generate consequent problems for future administration and delivery. It is therefore probably not a fully viable solution to employ these mechanisms beyond the narrow, project-specific confines for which they were designed.

Serializing CERIF into the METS structural map

A third strategy for integrating CERIF into METS attempts to create a logical structural map which encodes the structure of the complex web of relationships embedded within the CERIF metadata itself. This is undoubtedly a non-standard use of the METS schema: the structural map is designed to encode relationships between the component data files that make up a digital object, not the metadata associated with it. Nonetheless, a case may be made for such complex and fragmented metadata as CERIF to be handled within METS in this way: the implications for the standard will be discussed after the methodology itself is explained.

Under this approach, a series of structural maps are constructed in which the internal hierarchy of elements serializes the connections expressed in CERIF's linking files. A 'striped' syntax (as advocated by Habing & Cole (2008) for expressing aggregations) is here used to serialize the required relationships. A structural map for a typical CERIF linkage (once more between an author and a research output) takes the following form:-



The linkage itself is expressed in three nesting elements (marked by a in the diagram): the outermost delineates the first component (here the author identifier) of the link, using the method of an xpointer in the BEGIN attribute of an <area> element which references the element with the embedded CERIF file containing this information. The next level of the hierarchy indicates the semantic term used to characterise the link, again using the xpointer syntax to point to the element in CERIF's class definition file in which this term is declared. The lowest level of the hierarchy contains the object of the linkage, in this case details of the research publication. Although the nesting itself is sufficient to designate the function of each within this hierarchy, the TYPE attribute of may be used as in this example to render the function more explicit by employing a controlled set of terms (here "link-subject", "cfClassId" and "link-object"). Similarly, although not necessary, the LABEL attribute may be used as here to record the

value of each referenced CERIF element, in order to render the structural map itself clearer to the reader.

Two further elements, marked by b in the diagram, contain information on the temporal limits of this linkage which are an obligatory part of the CERIF model. Here the xpointers reference the link table itself, the only part of the CERIF data set in which this information is contained. It will be noticed that one further element in the linking table, <cfClassSchemeId>, the identifier for the classification scheme from which the semantic element <cfClassId> is taken, is not encoded in the hierarchy: this, however, is unambiguously derivable from the CERIF class file as the following sibling of the <cfClassId> element referenced by the second in hierarchy a.
A similar approach may also be used to handle such features as CERIF's tables for encoding multiple-language versions of textual data (such as titles or abstracts of bibliographic works) as shown in figure 3 below. Encoding relationships of this type using ´striped´ elements is relatively

simple:-

In this case only two levels of hierarchy are necessary to express the relationship between a component and its multi-lingual manifestations. Generating multiple language views of the CERIF object is then a simple matter of using XSLT transformations with a parameter for desired language codes to select those components with matching cfLangCode attributes.

Serializing CERIF relationships in this way does appear to combine the best parts of both of the two approaches discussed earlier. It employs the CERIF standard itself, rather than a bespoke intermediary, and also renders the resolution of the linkages inherent in CERIF much easier to handle in a working environment, and easier to document (for instance, in a METS profile) than raw CERIF itself may prove. It is, therefore, likely to be easier to exchange METS files with CERIF encoded metadata using this technique than would be possible using the first two methods.

Using the structural map in this way may, however, legitimately be seen an inappropriate use of the METS schema as it is currently constructed. The structural map is not specifically designed as a mechanism for describing relationships within metadata, and it is necessary to employ something of a fudge (in the form of recursive referencing) to make it work here. It may also be argued with some validity that the structural map is not designed to allow the semantic modelling of the type envisaged here, and to do so severely stretches the METS model beyond its intended use.

The latter argument may reasonably be countered by arguing that the METS structural map has often been used for encoding logical as well as physical structures in which, it could be argued, some degree of semantic definition is inherent (even in such basic notions as 'chapter' or 'section' in a book, for example). It could be argued that the approach taken here defines logical structures in an analogous way, albeit one that extends the concept of a logical structural map beyond that originally envisaged.

The arguments against the use of the structural map to serialize metadata relationships are certainly more serious. This approach undoubtedly extends the METS's functionality into areas for which it was not originally designed, but a case could be made, in applications such as this where the metadata is highly complex and fragmented, for allowing some of the powerful structural features METS provides for data to be extended to metadata. As alluded to earlier, METS presently takes a rather monolithic 'bucket'-like approach to metadata (METS Editorial Board 2010, p.11) each bucket of which can only be referenced as a whole (through the DMDID attribute of for instance). In cases such as this, where more sophisticated referencing would be useful, the standard could usefully be extended to allow it to address individual metadata components as envisaged here. A change of this kind would obviate the need to take a non-standard approach of this kind.

Conclusions

Providing contextual information on the research environment surrounding a publication or output is becoming increasing useful, and so employing a methodology which allows a key digital object metadata packaging scheme to be used with the only established research information standard currently available is a sensible option. The complexity of the CERIF standard, however, does present a number of problems for integrating it with METS.

Although it is possible to employ CERIF as a standard extension schema, or to employ an 'intermediary' schema instead, there is a case for serializing within METS itself some of the CERIF relationships which require decoding to make sense of the metadata expressed with it. At present METS does not readily allow this treatment of metadata, and so a non-standard methodology is required to allow it to be employed. There is a possible case for amending the METS standard to allow this type of metadata handling, although this would require extensive further investigation. Integrating CERIF into METS, however, is undoubtedly an important way to an integrated approach to digital library metadata to incorporate complex contextual information of this type without abandoning its core architectural principles.

References

Bolton, S. (2010). The Business Case for the Adoption of a UK Standard for Research Information Interchange. Joint Information Systems Committee. Retrieved from http://www.jisc.ac.uk/publications/reports/2010/businesscasefinalreport.aspx

Centre for e-Research (2011) 'R4R: Readiness for REF', R4R: Readiness for REF, [online] Available from: http://r4r.cerch.kcl.ac.uk/ (Accessed 7 February 2011).

Consultative Committee for Space Data Systems. (2009). Reference Model for an Open Archival Information System (OAIS): Draft Recommended Standard. Retrieved from http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf

Digital Library Federation. (2010). <METS>: Metadata Encoding And Transmission Standard: Primer And Reference Manual. Digital Library Federation. Retrieved from http://www.loc.gov/standards/mets/METSPrimerRevised.pdf

European Organisation for International Research Information. (2010). CERIF Introduction. Retrieved February 8, 2011, from http://www.eurocris.org/Index.php?page=CERIFintroduction&t=1

Fedora Project. (2006). Encoding Fedora Objects in METS (Fedora Extension). Retrieved April 19, 2011, from http://fedora-commons.org/download/2.1.1/userdocs/digitalobjects/rulesForMETS.html

Gartner, R. (2008). Metadata for digital libraries: state of the art and future directions. JISC Technology and Standards Watch. JISC. Retrieved from http://www.jisc.ac.uk/media/documents/techwatch/tsw_0801pdf.pdf

Gartner, R. (2011) 'Intermediary schemas for complex XML publications: an example from research information management', Journal of Digital Information, 12(3), [online] Available from: https://journals.tdl.org/jodi/article/view/2069 (Accessed 7 July 2001).

Guenther, R., & Myrick, L. (2006). Archiving Web Sites for Preservation and Access: MODS, METS and MINERVA. Journal of Archival Organization, 4(1-2), 141-166.

Habing, T., & Cole, T. (2008). ORE Resource Map Implementation in METS: draft proposal. Retrieved April 1, 2011, from http://ratri.grainger.uiuc.edu/oremets/1structMap_1REM.htm

Higher Education Funding Council for England. (2007). Data collection user guide: how to input, manage and submit data for the RAE. HEFCE. Retrieved from http://www.rae.ac.uk/pubs/2007/01/userguide/RAE%20User%20Guide.pdf

Jeffery, K. G., Lopatenko, A., & Asserson, A. (2002). Comparative study of metadata for scientific information: the place of CERIF in CRISs and scientific repositories. Gaining Insight from Research Information (pp. 77-87). Presented at the Sixth International Conference on Current Research Information Systems, University of Kassel: Kassel University Press.

Library of Congress. (2010). Metadata Object Description Schema: MODS. Retrieved January 28, 2010, from http://www.loc.gov/standards/mods/

Library of Congress. (2011a). Metadata Encoding and Transmission Standard (METS) Official Web
Site. Retrieved March 4, 2011, from http://www.loc.gov/standards/mets/

Library of Congress. (2011b). PREMIS: Preservation Metadata Maintenance Activity (Library of Congress). Retrieved January 28, 2010, from http://www.loc.gov/standards/premis/

Massachusetts Institute of Technology. (2010). DSpaceMETSSIPProfile - DSpace - DuraSpace Wiki. Retrieved April 19, 2011, from https://wiki.duraspace.org/display/DSPACE/DSpaceMETSSIPProfile

Maslov, A. et al. (2010) 'Adding OAI-ORE Support to Repository Platforms', Journal of Digital, 11(1), [online] Available from: http://journals.tdl.org/jodi/article/view/749/640 (Accessed 31 May 2011).

METS Editorial Board. (2010). Reimagining METS: an exploration: draft, draft, draft for discussion at Fall 2010 DLF. Retrieved from https://docs.google.com/fileview?id=0BzMhAsKzul-tN2Y0N2RjM2EtNzFmMC00NGQ3LTkyYjctNGZlZjEwODUzNmFm&hl=en&authkey=CIWlms8F

Rogers, N., Huxley, L., & Ferguson, N. (2010). Exchanging Research Information in the UK. Retrieved from http://ie-repository.jisc.ac.uk/448/1/exri_final_v2.pdf

University of Southampton. (2010). Preservation Support - EPrints. Retrieved April 19, 2011, from http://wiki.eprints.org/w/Preservation_Support

Article 4

Intermediary schemas and semantic linkages: an integrated architecture for complex digital archives

International Journal of Metadata, Semantics and Ontologies. 9 (4), 289-298.

This is an electronic version of an article published in International Journal of Metadata, Semantics and Ontologies, 9(4), pp. 289–298. The article is available online at: https://doi.org/10.1504/IJMSO.2014.065437

Intermediary schemas and semantic linkages: an integrated architecture for complex digital archives

Abstract

This article attempts to assess the feasibility of a Metadata Encoding and Transmission Standard (METS) based XML approach to integrated metadata for a complex digital archive. In particular, it aims to test whether such an approach can emulate two key features of RDF-based metadata architectures: the flexible reusability of components and the encoding of semantic linkages. In doing so, it seeks to establish whether this approach can be a viable alternative to ontology-based models for digital archive metadata. To do this, the standard use of METS as a packaging schema is extended as an 'intermediary schema' to enable the reuse of conceptual models within its architecture; in addition, the semantic mapping of components to concept repositories is achieved using the METS structural map and the Metadata Authority Description
Schema (MADS) schema for controlled vocabularies.

Keywords

metadata; XML; intermediary XML schemas; METS; MADS; semantic linking; digital archives; conceptual models; metadata reuse; ontologies.

Introduction

Despite great advances in search and retrieval technologies made over the last decade, exemplified perhaps by the growth of Google and other search engines, the need for digital collections to implement a coherent metadata strategy remains acute. Metadata remains essential to allow digital collections to be found, accessed, maintained, preserved and used intelligently.

For a metadata strategy to be effective in the management and presentation of complex objects composed of disparate media requires an approach which treats the entire information architecture as an integrated whole. Individual components of this architecture, such as the many standards which already exist for descriptive, administrative and technical metadata, need to be linked to each other in consistent and coherent ways if collections of any degree of size or complexity are to operate as effectively as the digital medium potentially allows. This need becomes particularly acute as collections and their component objects are linked to others within federated environments.

Two approaches offer viable methodologies for integrating digital collections in this way. One, exemplified by the Metadata Encoding and Transmission Standard (METS) schema (Library of Congress, 2011b), attempts to 'package' metadata components within eXtensible Markup Language (XML) architectures (McDonough, 2006).

The other employs the semantic web technologies of 'linked data', usually encoded within Resource Description Framework (RDF) (World Wide Web Consortium, 2004a) architectures, to join atomistic metadata components into integrated frameworks. The popular Fedora Commons digital repository system, for instance, makes extensive use of this model for its underlying architecture (Lagoze et al., 2006).

Although the semantic web approach to integration is gaining widespread attention, the use of XML architectures as integrative mechanisms has been receiving less advocacy recently. The purpose of the research underlying this article is to test the feasibility of this approach in the context of a complex digital archival environment that makes demands, such as semantic interoperability, which are often claimed to be the domain of semantic web methodologies.

By establishing the viability of integrated XML architectures for information environments of this type, it also aims to investigate its potential advantages over RDF-based models. Following an initial discussion of the concepts underlying both approaches, their perceived advantages and disadvantages and a brief survey from the literature of recent research on both techniques, the article introduces the data archive which forms the basis of this study and its demanding metadata requirements. The overall XML architecture created to meet these needs is then discussed, concentrating particularly on how it emulates two features often claimed as core advantages of linked data, the reusability of atomistic components and the facility to make semantic linkages at all levels of granularity. A further section then discusses how these architectures can be rendered in a working information environment (in this case using Fedora Commons). A final section analyses the results of these findings in the light of current research.

Background: approaches to integrated metadata

Metadata, and hence a coherent metadata strategy, remains essential for digital archives to operate effectively despite the increasing power and intelligence of search engines. As Smith (2004) points out: "although fully searchable text could, in theory, be retrieved without much metadata in the future, it is hard to imagine how a complex or multimedia digital object that goes into storage of any kind could ever survive, let alone be discovered and used, if it were not accompanied by good metadata". Allowing digital objects to be interoperable within an archive and enabling the archives themselves to interoperate through federated searching and browsing requires an integrated approach to the metadata environment.

Allowing semantic interoperability is an essential feature of such an environment. The definition of semantic interoperability is inevitably contentious, although a particularly useful typology is that adopted by Pollock and Hodgson (2004, pp.259–346) which specifies seven types: interoperability of data, process, services, taxonomy, applications, policies and social networks. In addition, it is also useful to consider several layers of interoperability, such as the four (structural, semantic, record and repository levels) suggested by Zeng and Chan (2010, p.4650). Not all approaches to metadata integration and semantic interoperability may necessarily be able to address every type and level, although the metadata environment as a whole for an extensive archive should attempt to do so.

In the analogue library environment, the advent of the standards MAchine-Readable Cataloguing (MARC) (Library of Congress, 2013a) and *Anglo-American Cataloging Rules*, 2nd ed. (AACR2) (American Library Association, 2012) facilitated this integration, allowing the large union catalogues which are a core part of the library world to be constructed. These standards, and the intended successor to AACR2, Resource Description and Access (RDA) (Library of Congress, 2013b) concentrate primarily on descriptive metadata: the digital environment also requires an extensive set of technical, administrative and structural metadata to support the delivery and curation of digital objects. To produce an integrated environment for this metadata requires a method of linking these disparate metadata types into coherent architectures. Several approaches have been advocated for facilitating this.

One well-established approach relies on the structures of XML schemas as a 'packaging' architecture. These schemas, constructed using the W3C XML Schema Definition Language (World Wide Web Consortium, 2004b), generally prescribe strict sets of nested hierarchies descending in nodes from a top-level root element. This tends to be 'XML as metadata' as defined by Salminen and Tompa, i.e. semantic information inherent in the schema definition itself (Salminen and Tompa, 2011, p.150). The arrangement of these components within such schemas is analogous to the strict hierarchies of an enumerative taxonomy, although these structures may be obviated by the use of linking mechanisms which cut across their boundaries.

An alternative to this is to use semantic web methodologies, usually encoded in a form of RDF-linked data, to facilitate integration of this type. This approach atomises metadata (and data) components to fine levels of granularity and then builds architectures which link these by defining RDF 'triples', subject–predicate–object linkages, which together form an integrated network of semantically joined units of information. Unlike the hierarchical, tree-like structures preferred by XML schema definitions, RDF attempts to build a graph-like network of interconnected semantic units. To repeat the analogy with taxonomic systems, this approach is much closer to the faceted than enumerative approach, offering much greater flexibility in the use and combination of components.

The former approach is exemplified by a report published in 2008 by UK's Joint Information Systems Committee (JISC) (Gartner, 2008). This document advocates a strategy based on a combination of existing XML standards for descriptive, administrative and technical metadata integrated within the METS schema. It claims that by employing metadata schemes which adopt Functional Requirements for Bibliographical Records (FRBR) as their underlying conceptual model within the shared syntactical framework of XML, a coherent sector-wide metadata landscape can be constructed (Gartner, 2008, p.6). It is also employed to some degree by the Darwin Information Typing Architecture (OASIS, 2013) which uses XML schemas to create and publish information at flexible levels of granularity.

The use of hierarchical XML approaches to integration has extensive advocacy in the digital library community, for such reasons as its platform independence, extensibility, modularity and interoperability (McDonough, 2008) and, with its predecessor SGML, has for many years been recognised as one of the most archivally robust of

formats for digital library metadata (Coleman and Willis, 1997). The widespread adoption of key XML-based standards such as METS testifies to the practicality of implementing XML in working environments.

Nonetheless, doubts have been expressed about its ability to allow semantic in addition to *Intermediary schemas and semantic linkages* 291 syntactic interoperability (Gradmann, 2010, p.159) and some authors have questioned even its ability to allow the syntactical interoperability of structural metadata (McDonough, 2008). The relative hierarchical rigidity of XML architectures has also been criticised for impeding users' discovery options and limiting administrators' management of resources (for instance, in extending hierarchies to fit local metadata requirements (Han, 2006, p.236)).

The use of RDF graph-based architectures has been adopted by such key open-source repository systems as Fedora Commons, and forms the basis of a number of recent projects, such as the Biophysical Repositories in the Lab (BRIL) project at King's College London (http://bril.cerch. kcl.ac.uk/). Two key benefits are often claimed for the use of these over the XML approach which is currently better established within the library community: these are the potential it offers for the flexible reusability of components (Lagoze et al., 2006, p.127) and its ability to encode complex semantic relationships between objects at multiple levels of granularity (Lagoze et al., 2006, p.135). Further advantages have also been enumerated for the RDF model in allowing mappings flexible enough to incorporate new conceptual models for bibliographic data such as the Functional Requirements for Subject Authority Data (FRSAD) (International Federation of Library Associations, 2013) scheme (Zeng and Zumer, 2009, pp.39–44).These advantages are undoubtedly valid and potentially very powerful in the context of complex data and metadata environments. In working digital library environments, however, they have yet to fully manifest themselves to a major extent, and some scepticism about the utility of this approach is often expressed. A recent report for JISC, for instance, concluded that "we have yet to see any real examples of benefit [from linked data for library metadata] emerging from JISC projects in this area, or elsewhere" (Hawtin et al., 2011, p.26) and a major UK-based digital library vendor, Talis, has recently withdrawn from semantic web operations (Anon, 2012).

The most problematic disadvantage of this methodology at present is that it remains a highly complex way of encoding even simple content models, owing to the graph-based architecture of RDF: this can have major implications for processing and query times. In particular, it has been noted that data modelling and ontology development can be very time-consuming, data cleansing is often problematic (Hawtin et al., 2011, p.17) and that the particular demandslinked data makes on developers can lead to skills shortages (Hawtin et al., 2011, p.30). It has also proved more difficult to exchange and reuse content models than the theoretical interoperability of RDF might suggest (Sharma, 2007).

Some recent research, which extends the already long established body of research on the use of XML for digital library metadata, has endorsed the use of the XMLpackaging model advocated by Gartner (2008). Its espousal, for instance, of the XML syntax as central to metadata integration and interoperability has been cited as key to handling compound digital objects (Dulock and Cronin, 2009, p.299) and its use of a collection of

standards united by a common conceptual model has been cited as an important basis for digital preservation in libraries (Chen and Reilly, 2011, p.86). Recent work has also advocated a similar METS-based approach for cloud-based digital archives (Askhoj et al., 2011).

None of this research, however, examines the questions of the flexible reuse of component features and semantic interoperability which are often cited as the main strengths of the alternatives. The purpose of this study is to establish whether mechanisms can be used within XML architectures to emulate some of the beneficial features of the semantic web environment: specifically, it aims to find methodologies for encoding semantic relationships between objects at multiple levels of granularity and to allow the reusability of features of components in a manner analogous to that of Fedora Content Models. In this case, the use of XML moves from 'XML as metadata' to 'XML for metadata', to employ Salminen and Tompa's distinction once more: the semantics of the XML schema here become mechanisms to support external semantics which form the essential definition of an application.

Although it remains an open question somewhat beyond the scope of this article whether all structural metadata expressible in hierarchical XML structures can also be expressed in RDF and vice versa, the features necessitated by the particular requirements of the data archive described below have potentially wide applicability in complex environments where the atomistic flexibility of semantic web methods may initially seem attractive.

The Demonstration Test Catchments Archive

As a case study for the techniques devised here, a digital archive of environmental and associated data collected as part of the UK Freshwater Biological Association's Demonstration Test Catchments (DTC) project (http://www.demonstratingcatchmentmanagement.net/), was chosen. Catchments are drainage basins where surface water enters a river and as such provide sources of key data on pollution for complete ecosystems. The DTC project collects data on water quality from the catchments of three rivers in England, the Eden, Wensum and Avon, to investigate the effects of pollution from agriculture on ecosystems. The data are relevant to the work of scientists, farmers, environmentalists and policy makers, all of whom could benefit from the detailed evidence base that the project aims to collect.

The diverse metadata requirements of the research community served by this archive have already been examined in the literature on the recently completed FISHNet project (Hedges et al., 2012). The types of data accommodated in the archive are diverse; in addition to document objects (such as reports, publications and emails), the collection includes photographs, an extensive variety of spatial data and time series data gathered from the catchments themselves: these are often small, 'hand crafted' data sets, each of which uses its own approach to data and metadata.

Two features of this archive's metadata requirements make it suitable for testing methodologies for reusing components and embedding semantic linking. Reusing components in a flexible and multiple-granular way is

required because of the archive's diverse range of both depositors and users and their differing requirements for access to the data. The archive has identified an extensive list of 'personae', categories of users with different expectations, skills and intended uses for the archived data. Both input and access mechanisms for the metadata for each data type need to be tailored to the requirements of these user types. These will vary greatly according to the data types themselves and also the requirements and skills of the user: the same data set, for instance, will be accessed in very different ways by a scientist with proficient data analysis skills and a local farmer whose information technology knowledge may be much more limited.

Semantic interoperability is required as all data at any level of granularity within the digital objects deposited should be available for search, extraction and analysis in a federated manner. It should, for instance, be possible to extract data on the concentration levels for any chemical wherever they are found in the objects within the archive, and however they are described within each object: for example, it should be possible to combine data on chemical traces in water samples across all measurement stations to produce visualisations of variations in levels. To achieve this, semantic equivalences between components need to be recorded, so that the same conceptual object, however it is labelled within the archived data object, can be co-identified with its other manifestations wherever they are located.

Reusing metadata components: conceptual models for data types and personae

To encode the key features of each data type and persona, it is necessary to define a series of 'conceptual models'. These attempt to emulate the reusability of RDF-based metadata units within the XML syntactical model by acting as templates for the metadata specifications for each data type in a manner analogous to the content model approach of Fedora. These models may include, for instance, encoding mechanisms for metadata entry or mechanisms for its delivery to different categories of users.

Implementing these models requires the use of 'intermediary' XML schemas (Gartner, 2011, 2012). The principle of intermediary schemas as advocated by Gartner (2011) is that XML schemas can be used either to constrain highly flexible metadata schemes with complex architectures (2011) or to provide templates which can alleviate the problems of handling metadata for complex multi-media objects by separating the conceptual models underlying their internal structures from the metadata itself (2012). These schemas are therefore designed to act not as final delivery containers for metadata but as mediating mechanisms from which their final form is generated by XSLT transformations (Gartner 2012, p.29).

In Gartner (2012), this technique is used to handle variable requirements for structuring data for user delivery, by encoding templates at any level of granularity from which the often large and complex XML files for final delivery can be constructed on-the-fly. In this way, for instance, the sequence of still images needed for a video can be handled by separating the metadata for the images themselves from the metadata for their sequencing and the overall metadata architecture for the complex object in which they are delivered (Gartner 2012, p.32). This technique offers great potential for enhancing the flexibility of otherwise rigid XML architectures. In Gartner (2012), it is used exclusively as a method of facilitating the delivery of complex objects, but here it is

extended to facilitate the diverse requirements for data entry necessitated by a heterogeneous community of data providers. To do this, the same methodology of using XML to encode conceptual models or templates is employed, but in this case it is employed to associate their use with 'personae' (defined user groups).

The packaging mechanism employed in this archive is the "Metadata Encoding and Transmission Standard" (METS) (Library of Congress, 2011b) schema. METS is designed to allow all metadata for a digital object, descriptive, administrative and structural, to be collocated in a single extensible XML architecture. Its core component is a hierarchical structural map used to express the relationships between an object's internal components, each of which may be linked by XML identifiers to any piece of descriptive or administrative metadata at multiple levels of granularity.

The final METS file which is used for this purpose is constructed by processing its component 'conceptual models' as required by the specifics of the data type used, or the user accessing it. To enable this to occur, each conceptual model must be defined, and each data type or persona must be associated with the respective model(s) which will be activated when it is delivered. For both of these functions, METS itself is used as an intermediary schema, defining small templates which are combined to produce the larger METS file which is used for delivery.

As an example of this, the intermediary schema mechanism may be used to associate each data type with an XFORM file. XFORM (World Wide Web Consortium, 2009) is an XML standard for encoding reusable and interoperable forms for data entry: associating an XFORM in this way ensures a consistent, archive-wide approach to data entry which can readily incorporate new data types. In addition to this mechanism for data entry, the intermediary schema technique can rationalise the delivery of objects by associating multiple XSLT transformations with each data type, each designed for a specific class of user (or persona). Both of these types of association are readily handled by encoding them in a template METS file, as is shown in Figure 1.

Figure 1 Incorporating XFORM and XSLT associations into template METS file

Each data type is assigned a separate element within a hierarchy of types in the METS structural map (labelled a in the diagram) and, following standard METS practice, each is assigned a unique ID. Each also contains an <mptr> element, designed to reference a further METS file containing the content of a given METS component, but here acting as a placeholder to mark where this content should be generated when the file is processed as an intermediary schema. This extends the technique advocated in Gartner (2012) by allowing the recursive use of <mptr> elements within the intermediary schema structure, thus enabling greater levels of complexity to be handled within this single template architecture.

Each XFORM or XSLT is listed in a separate <behaviour> element (labelled b and c) within the behaviour section, the location in an METS file for associating mechanisms for rendering or processing a digital object. The BTYPE (behaviour type) attribute here indicates the function of the mechanism (here set to 'INPUT' to designate the XFORM for metadata input and 'PERSONA' to indicate an XSLT transformation for a given user category). In the case of the XSLT transformations, the user category for which each is intended is listed in the GROUPID attribute, which can contain any value required by the metadata provider to subdivide mechanisms by function: in the example shown, this is set to 'Public' to indicate that this XSLT transformation renders the file for use by the general public.

Each behaviour is then linked to the divisions in the structural map to which it applies by its STRUCTID attribute which contains a reference to the ID of the element for its respective data type; in the example shown, both are linked to the division for flow gauge data, and so are only activated for that data type.

This mechanism allows small atomistic metadata components to be combined within an overall METS architecture in a flexible and reusable way, thus emulating some of the functionality of RDF content models but retaining the advantages highlighted above of the robust XML architecture of METS.

Semantic linking of object components

The semantic linking of data components at any level of granularity is necessitated by the requirement for the archive to function as an integrated tool for data analysis. This is readily achieved using semantic web techniques, and is claimed to be the key advantages of RDF over XML. In the METS-based XML environment described here, this is achieved by establishing an architecture for mapping metadata components to a central, archive-wide, vocabulary of concepts.

In the example shown in Figure 2, water flow velocity in a spreadsheet (marked Vel.) is mapped to a table in an mySql database in which it more explicitly labelled 'WaterVelocity'. To achieve this mapping, the columns for water flow velocity in the spreadsheet and database are linked to a central registry of concepts and from there to each other; it is then possible to co-analyse data in diversely located files (such as the two numerical entries linked in the diagram).

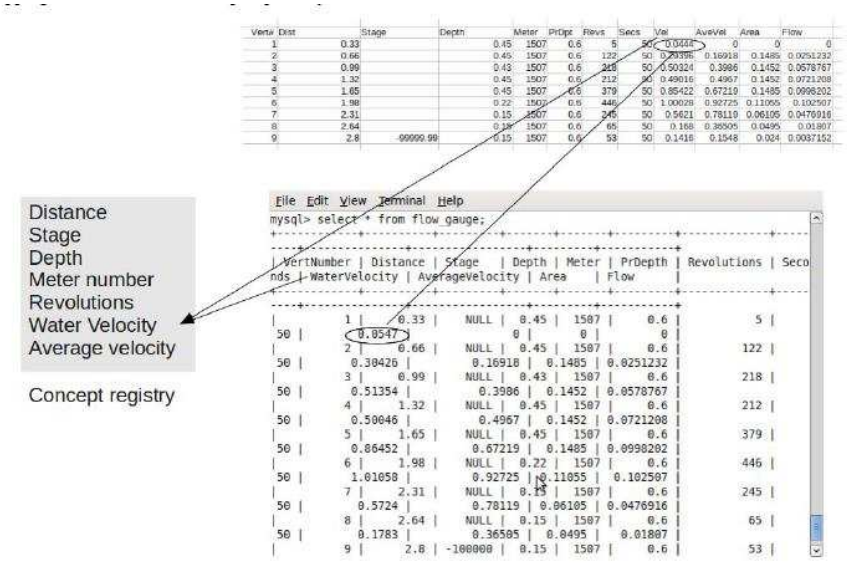


Figure 2 Mapping data entries via a concept repository

To achieve this in RDF would involve mapping instance data from the source files to an ontology containing the central concept vocabulary and their relations. The ontology itself would be relatively straightforward to implement using Web Ontology Language (OWL) (World Wide Web Consortium W3C, n.d.). Generating RDF triples from the source spreadsheets would involve extracting each data value as a literal and incorporating it into a triple as an object. This would best be done using a script written in Perl or Python to extract the data from CSV files converted from the spreadsheets and to generate triples linking these literals to the URIs of concepts held in the ontology using a suitable predicate (such as *owl:hasValue*).

Such a process could be presented diagrammatically in this way.

The triple for this single data would have to be supplemented by a series of others to establish the contextual environment in which it is present. It would require links, for instance, to other components on the same row, others links to identify the spreadsheet in which it is located, yet others to establish its data type in order to enable its processing and so on. As mentioned earlier, a complex set of atomistic triples is soon built up, which brings with it associated problems of maintenance and preservation.

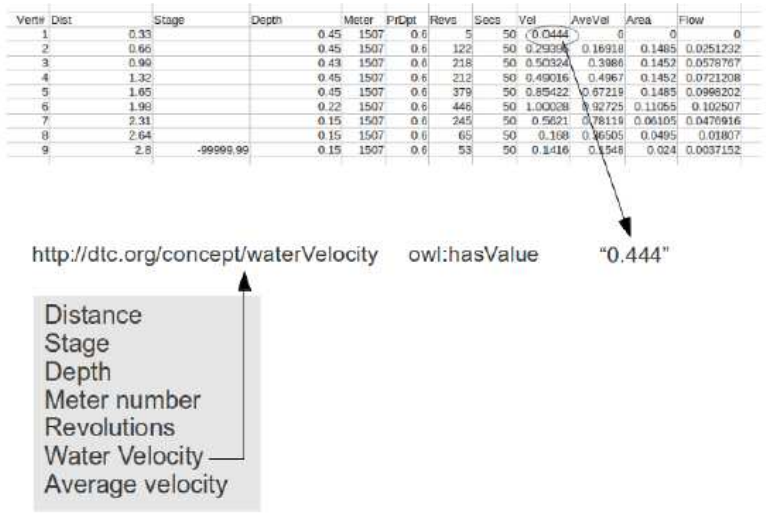


Figure 3 Encoding data component linkages in RDF

In the alternative approach advocated here, these linkages are instead encoded using <area> elements within the METS structural map, and a single controlled vocabulary file for the archive as a whole is encoded in Metadata Authority Description Schema (MADS) (Library of Congress, 2011a). MADS is an XML schema designed by the Library of Congress specifically for the encoding of machine-readable vocabularies and thesauri, and allows interoperable and transferable authority lists to be constructed and shared in a systematic way. MADS is published primarily as an XML schema, although it is also available as an RDF ontology. Both follow the practice advocated by developments such as FRSAD in separating concepts from their 'labels' (Zeng et al., 2012, p.107), thus allowing much greater flexibility in the handling of authority control than has previously been possible. It is a companion to the widely used Metadata Object Description Schema (MODS) (Library of Congress, 2010) schema for descriptive metadata, with which it integrates seamlessly.

The overall architecture for this structure of linkages is shown for some sample data in Figure 3.

The <area> element within divisions of the METS structural map is designed to allow the precise addressing of individual sections or components of the files from which the digital object is composed: this mechanism may be used, for instance, to subdivide a video file by time codes or to highlight a part of a digital image by pixel coordinates. For data files, byte offsets are one mechanism by which these components can be addressed: in the examples shown in the diagram, Comma-Separated Value (CSV) files are generated from a spreadsheet and

mySql database, and the BEGIN and END attributes of the <area> element are used to record the byte offsets of each value. These <area> elements are then grouped within the structural map by the concept whose value they record: all water flow velocity components, for example, are grouped together by nesting them within the same . Each of these container elements, identified as a grouping of conceptual entities by its TYPE attribute being set to a value of 'DTC ARCHIVE CONCEPT', is mapped its equivalent concept in the central repository by the inclusion of a unique URI for the concept in its CONTENTSID attribute.

The central repository of concepts to which components are mapped by this mechanism is encoded in MADS. A key feature of this schema is that it allows each concept to be uniquely identified by URIs: in Figure 4, the sample MADS entries shown in the central box each contain a URI of this type which is recorded in the valueURI attribute. It is this URI, referenced from the CONTENTSIDS attribute of elements in the METS structural map for each object, which allows disparate manifestations of the same concept to be linked together.

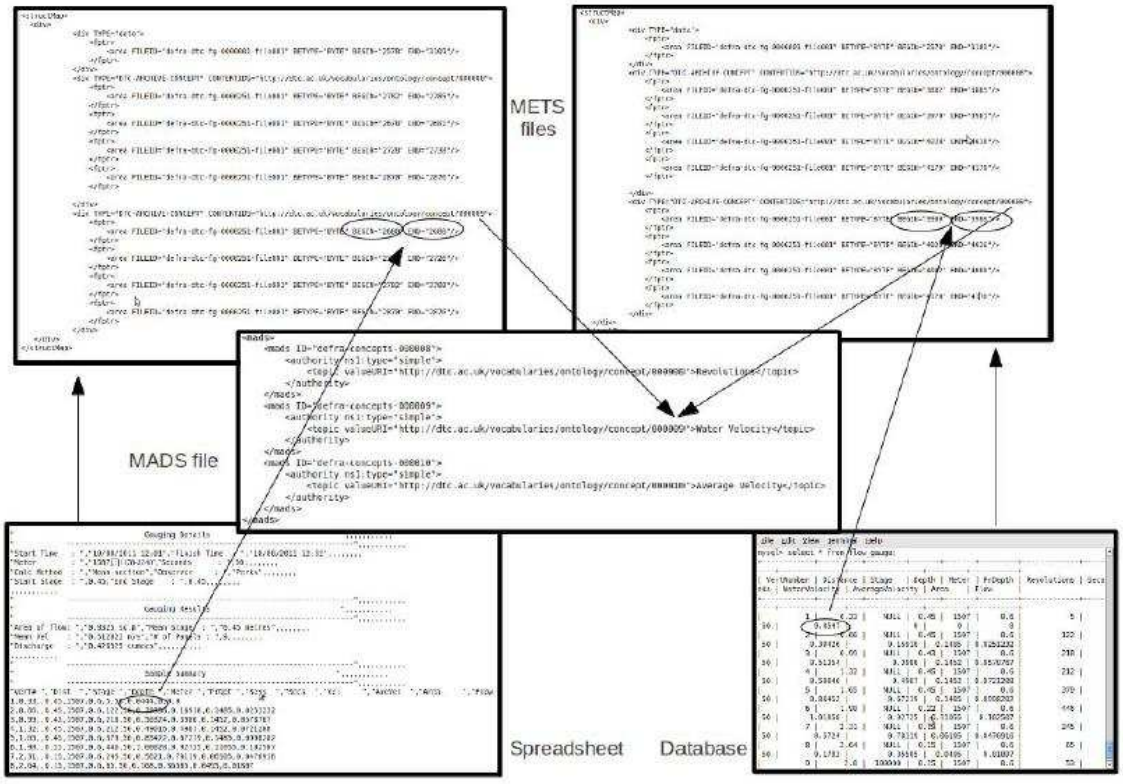


Figure 4 Encoding data component linkages with METS and MADS

This methodology allows the mapping of semantically equivalent components across the entire archive within an XML architecture. Extracting data for analysis across objects is a relatively simple process which may be carried out in a number of ways. Simple XPath statements, for instance, can be used in an XML search engine such as Lucene on-the-fly to extract data for processing. Although the processes involved in generating these semantic structures are relatively similar in complexity whether RDF or XML is used, the technique advocated here allows the full functionality of RDF linkages within the more easily managed environment of XML.

Rendering the objects

Rendering the objects for delivery entails using XSLT transformations to populate the template METS files for objects and conceptual models. Producing the final version of each digital object file is done by merging its METS file with the sections of the conceptual model file linked to the model to which it conforms. All sections of the file linked to this for the model are merged with all sections from the object's own METS file to produce the final version: this is then rendered as required to HTML or any other required delivery format.

Both of these processes are relatively simple transformations and can be done rapidly on-the-fly using XSLT engines such as Saxon. In the case of a Fedora archive, they can be speedily carried out by using the XSLT file as a disseminator for the XML stored as data streams.

Implications and conclusions

The techniques developed in the course of constructing this metadata strategy appear to emulate within the hierarchical structures of the XML architecture the two components of semantic web functionality cited above, the flexible reusability of components and the establishment of semantic linkages at any level of granularity. In doing so, they confirm the findings of Dulock and Cronin (2009), Chen and Reilly (2011) and Askhoj et al. (2011) that the XML packaging model can produce well-integrated metadata environments in working systems, but extend it further to incorporate these extended functionalities not usually associated with this methodology.

The reusability of components is achieved by the use of METS as an intermediary schema for collection-level and conceptual model descriptions; this technique allows the use of XML instances as templates or models which are realised only on delivery of the digital object. In this way, components are readily reused with the flexibility usually claimed for ontologically structured approaches to metadata (for instance by Lagoze et al., 2006) while retaining the advantages of the clear, integrated structures offered by XML. It also introduces a degree of flexibility into hierarchical XML architectures which to some extent obviates Han's (2006, p.236) concerns that these impede users and administrators from engaging with digital objects.

Semantic linkages at multiple levels of granularity are readily achieved by referencing entries in a central concept repository with the fine-grained component addressing features of the METS <area> element. Supplementing the METS architecture in this way with a MADS-based central collection of concepts, used for mapping widely spread data components together, allows the functionality of an ontological approach to linking data to function within the same XML framework. The MADS standard itself, designed specifically to encode controlled vocabularies of this type, is robust but still retains enough flexibility to meet the semantic and relational requirements of the vocabularies of concepts required for this functionality. Its underlying model should also ensure its compatibility with recent developments such as FRSAD. Gradmann's (2010, p.159) concerns about the semantic deficit of employing XML architecture appear to be met by this approach.

Using this technique requires the multiple processing of objects, once to generate the definitive METS object and once more to render it for delivery, but in the context of a high-performance system such as Fedora Commons, this will have a negligible effect on delivery performance. The same XSLT mechanisms and tools that are used to render and process the digital object and conceptual model metadata encoded in METS can be used to carry out this integration of data. There is therefore no need for the specialised and distinctive skills sets which Hawtin et al. (2011, p.30) see as an impediment to employing semantic web methodologies in working digital library environments, as the same, already established, methodologies for XML processing are employed to enable this functionality.

This integrated metadata architecture, which extends the methodology advocated by Gartner (2008) to new levels of granularity from collection (or super-collection) level down to individual data components, demonstrates that the XMLbased strategy advocated in that report can be applied to highly diverse and complex collections of resources. Its use of METS's methodology for integrating disparate metadata standards within its clearly delineated framework allows for a degree of structural integration at all levels; this reduces the complexity of delivering the archive's contents despite the often substantial disparities in their formats and contents.

The robust structure imposed by this strategy also ensures easy extensibility: new data types are readily incorporated by, for instance, new XFORMs or new XSLT transformations, all of which readily slot into the infrastructure and require little or no development work to enable their delivery to users. The techniques described here have some limitations of applicability and scope. They directly address, for instance, only two of Pollock and Hodgson's types of semantic interoperability outlined above, namely data and taxonomy, although they may also be useful in facilitating others if used as a base architecture and converted as necessary. They should, however, be applicable at all four of Zeng and Chan's levels of interoperability, although their focus is primarily on the semantic and structural. Further research could usefully be carried out of their applicability to other types of semantic interoperability, where the flexibility of RDF is perhaps particularly useful.

Using these techniques, it appears that XML architectures which share an underlying conceptual model can produce flexible, reusable metadata architectures even in the context of a highly diverse, evolving archive with multiple levels of granularity. If adopted more widely, the techniques presented here offer the potential for greater federated searching of collections which the semantic web technique, despite the potential of RDF for breaking down barriers between data systems, has only shown nascent signs of realising in working digital library environments.

References

American Library Association (2012) *Anglo-American Cataloging Rules homepage*. Available online at: http://www.aacr2.org/(accessed on 28 June 2013).

Anon (2012) *Talis shuts down semantic web operations*, Information Age. Available online at: http://www.

information-age.com/channels/information-management/news/2111803/talis-shuts-down-semantic-web-operations.html (accessed on 27 July 2012).

Askhoj, J., Sugimoto, S. and Nagamori, M. (2011) 'A metadata framework for cloud-based digital archives using METS with PREMIS', *Lecture Notes in Computer Science*, Vol. 7008, pp.118–127. Available online at: http://dx.doi.org/10.1007/978-3-642-24826-9_17 (accessed on 10 November 2011).

Chen, M. and Reilly, M. (2011) 'Implementing METS, MIX, and DC for sustaining digital preservation at the University of Houston Libraries', *Journal of Library Metadata*, Vol. 11, No. 2, pp.83–99.

Coleman, J. and Willis, D. (1997) *SGML as a Framework for Digital Preservation and Access*, Commission on Preservation and Access, Council on Library and Information Resources, Washington, DC. Available online at: http://www.clir.org/pubs/reports/pub68 (accessed on 4 June 2013).

Dulock, M. and Cronin, C. (2009) 'Providing metadata for compound digital objects: strategic planning for an institution's first use of METS, MODS, and MIX', *Journal of Library Metadata*, Vol. 9, Nos. 3–4, pp.289–304.

Gartner, R. (2008) *Metadata for digital libraries: state of the art and future directions*. Available online at: http://www.jisc.ac.uk/media/documents/techwatch/tsw_0801pdf.pdf (accessed on 23 August 2010).

Gartner, R. (2011) 'Intermediary schemas for complex XML publications: an example from research information management', *Journal of Digital Information*, Vol. 12, No. 3. Available online at: https://journals.tdl.org/jodi/article/view/2069 (accessed on 7 July 2012).

Gartner, R. (2012) 'METS as an 'intermediary' schema for a digital library of complex scientific multimedia', *Information Technology and Libraries*, Vol. 31, No. 3, pp.24–35. Available online at: http://ejournals.bc.edu/ojs/index.php/ital/article/view/1917 (accessed on 4 June 2013).

Gradmann, S. (2010) 'Digital Library Metadata', in Collier, M. (Ed.): *Business Planning for Digital Libraries: International Approaches*, Leuven University Press, Leuven, Belgium. pp.157–166.

Han, Y. (2006) 'A RDF-based digital library system', *Library Hi Tech*. Vol. 24, No. 2, pp.234–240.

Hawtin, R., Hammond, M., Miller, O. and Matthews, B. (2011) *Review of the Evidence for the Value of the 'Linked Data' Approach*, Final Report to JISC. Available online at: http://epubs.stfc.ac.uk/work-details?w=62217 (accessed on 27 July 2012).

Hedges, M., Haft, M. and Knight, G. (2012) 'FISHNet: encouraging data sharing and reuse in the freshwater science community', *Journal of Digital Information*, Vol. 13, No. 1. Available online at: http://journals.tdl.org/jodi/index.php/jodi/article/view/5884 (accessed on 27 November 2001).

International Federation of Library Associations (2013) *IFLA Working Group on Functional Requirements for Subject Authority Records (FRSAR)|IFLA*. Available online at: http://www.ifla.org/node/1297 (accessed on 1 August 2013).

Lagoze, C., Payette, S, Shin, E. and Wilper, V. (2006) 'Fedora: an architecture for complex objects and their relationships', *International Journal on Digital Libraries*, Vol. 6, No. 2, pp.124–138.

Library of Congress (2010) *Metadata Object Description Schema:MODS*. Available online at: http://www.loc.gov/standards/mods/ (accessed on 28 January 2010).

Library of Congress (2011a) *Metadata Authority Description Schema (MADS)*, Library of Congress. Available online at: http://www.loc.gov/standards/mads/ (accessed on 24 November 2011).

Library of Congress (2011b) *Metadata Encoding and Transmission Standard (METS)*, Official Web Site. Available online at: http://www.loc.gov/standards/mets/ (accessed on 4 March 2011).

Library of Congress (2013a) *MARC Standards*. Available online at: http://www.loc.gov/marc/ (accessed on 28 January 2010).

Library of Congress (2013b) *Resource Description and Access (RDA): information and resources in preparation for RDA*. Available online at: http://www.loc.gov/aba/rda/ (accessed on 28 June 2013).

McDonough, J. (2006) 'METS: standardized encoding for digital library objects', *International Journal on Digital Libraries*, Vol. 6, No. 2, pp.148–158.

McDonough, J. (2008) 'Structural metadata and the social limitation of interoperability: a sociotechnical view of XMLand digital library standards development', *Proceedings of the Balisage: the Markup Conference 2008*, 12–15 August, Montréal, Canada. Available online at: http://www.balisage.net/Proceedings/vol1/html/McDonough01/BalisageVol1-McDonough01.html (accessed on 10 July 2012).

OASIS (2013) *OASIS Darwin Information Typing Architecture (DITA) TC*, OASIS. Available online at: https://www.oasisopen.org/committees/tc_home.php?wg_abbrev=dita (accessed on 1 August 2013).

Pollock, J.T. and Hodgson, R. (2004) *Adaptive Information:Improving Business through Semantic Interoperability, Grid Computing, and Enterprise Integration*, Wiley-Interscience, Hoboken, NJ.

Salminen, A. and Tompa, F. (2011) *Communicating with XML*, Springer, New York.

Sharma, R. (2007) *Fedora interoperability review*. Available online at: http://wwwcache1.kcl.ac.uk/content/1/c6/04/55/46/fedora-report-v1.pdf (accessed on 15 June 2011).

Smith, A. (2004) 'Preservation', in Schreibman, S, Siemens, R. and Unsworth, J. (Eds): *A Companion to Digital Humanities*, Blackwell, Oxford. Available online at: http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-5-7 (accessed on 25 June 2012).

World Wide Web Consortium (n.d.) *OWL Web Ontology Language Overview*. Available online at: http://www.w3.org/TR/owl-features/ (accessed on 12 June 2012).

World Wide Web Consortium (2004a) *RDF – Semantic Web Standards.* Available online at: http://www.w3.org/RDF/(accessed on 28 June 2013).

World Wide Web Consortium (2004b) *W3C XML Schema.* Available online at: http://www.w3.org/XML/Schema (accessed on 2 July 2013).

World Wide Web Consortium (2009) *XForms 1.1*. Available online at: http://www.w3.org/TR/xforms/ (accessed on 8 December 2011).

Zeng, L.M. and Chan, L.M. (2010) 'Semantic interoperability', in: *Encyclopedia of Library and Information Sciences*, 3rd ed., M. Dekker, New York, pp.4645–4662.

Zeng, M.L. and Zumer, M. (2009) *Introducing FRSAD and mapping it with other models*. Available online at: http://www.slideshare.net/mzeng/introducing-frsad-andmapping-it-with-other-models (accessed on 1 July 2013).

Zeng, M.L., Salaba, A. and Zumer, M. (2012) *FRSAD: Conceptual Modeling of Aboutness*, ABC-CLIO, Santa Barbara, CA.

Article 5

An XML schema for enhancing the semantic interoperability of archival description.

Archival Science. 15 (3), 295-313

This is a post-peer-review, pre-copyedit version of an article published in Archival Science. The final authenticated version is available online at: https://doi.org/10.1007/s10502-014-9225-1.

An XML schema for enhancing the semantic interoperability of archival description

Abstract

The archival community is presently addressing questions of semantic interoperability, which have been raised by linked open data and the Semantic Web. Current discussions on revisions to encoded archival description (EAD), for instance, are moving from document-based to more database-like element sets to facilitate greater interoperability. But to realise fully the potential benefits offered by such developments, further steps are needed. This article outlines a new experimental schema, which allows collection descriptions to be encoded in a more interoperable manner. Devised for the Collaborative European Digital Archive Architecture Project, it complements the descriptive elements of a collection-level record with more semantically precise metadata components. The schema was devised to form the core of a 'digital ecosystem', a rapidly changing research environment of which archival descriptions form a core part. The advantages of semantic interoperability and the data requirements to achieve it are discussed before a detailed description of the schema itself. The article then suggests some initial ways in which it could be employed to enhance access to collection descriptions, its compatibility with the Semantic Web and standards such as ISAD(G), and how it may be used in conjunction with EAD. The schema has yet to be evaluated in detail in working environments, and so is offered as a heuristic proof-of-concept for community discussion.

Keywords

Metadata Archival description Metadata interoperability Encoded Archival Description Semantic Web

Introduction

In its 20 years of existence, the encoded archival description (EAD) (Library of Congress 2011) has become the effective *lingua franca* for machine-readable descriptions of archival collections and their contents. To a large extent, it has fulfilled its primary objective as defined by one of its authors, Daniel Pitti, to provide 'archivists and both professional and public researchers [with] universal, union access to primary resources' (Pitti 1999) in a manner akin to the federated access to bibliographical records facilitated by the MARC standard in libraries. Its standardised method of description has enabled services such as the Archives Hub (MIMAS n.d.) or AIM25 (King's College London Archives n.d.) to emulate to an extent the functionalities of large-scale union catalogues in the library sector.

Substantial as these achievements are, EAD has been subject to some criticism for a perceived failure to enable the degree of interoperability for archival records to which the online world aspires. Several reasons have been given for this, including, for instance, the flexibility of the EAD architecture, which allows similar concepts to be recorded in multiple ways (Shaw 2009, p. 123). Others, such as Dow (2009), argue that EAD is inherently more 'document-centric' (modelled on the requirements of encoding the features of the textual

document which forms the traditional printed finding aid) than 'data-centric' (modelled on atomistic data components at multiple levels of granularity); this makes interoperability, particularly its potential for detailed federated searching, harder to achieve.

Suggestions have been made in the literature in recent years for the directions in which archival description can move to realise more fully the potential of the machine-readable medium. Pitti himself has expressed aspirations for an environment which combines database and markup functions to provide a more 'complete and flexible system of archival description that would interrelate record description, creator description and the description of functions and activities' (personal correspondence, quoted in Shaw 2009, p. 113). Bunn also argues for moving away from archival descriptions towards metadata for records, perhaps developing standards such as ISO 23081 for this purpose (Bunn 2013).

This article outlines an experimental approach to archival collection description, which aims to extend the functionality of EAD to allow data and markup functions to be more readily combined in the manner proposed by Pitti. The centre of this approach is an XML schema devised as part of Collaborative European Digital Archive Architecture (CENDARI) (Cendari Project 2013), a European Commission-funded project that aims to provide a research infrastructure to integrate the digital collections of multiple archives and libraries in the subject areas of mediaeval and modern European history. To enable the degree of integration required for such an infrastructure to operate requires semantic interoperability at a fine level of granularity. This is achieved by incorporating some of the methodologies underlying the Semantic Web into the structured architectures of the XML environment and relating these to the requirements of EAD and archival description in general.

Interoperability and archival description

Interoperability between resources, the ability seamlessly to exchange and cross-search information, is one of the key benefits of machine-readable data and the rationale behind most metadata standards in the library and archival sciences. Such interoperability was a key motivation behind the creation of EAD from its inception: in the words of Pitti (1999), 'standardisation will support the long-cherished dream of providing archivists and both professional and public researchers universal, union access to primary resources'. Interoperability of this kind is easier to achieve the smaller and more precise the data units are which form the linkages between resources: it is, for instance, easier to establish that a date formatted in a precise way is equivalent to a similarly formatted date in another resource than to assert the equivalence of two textual descriptions.

The most widely discussed method for enabling such interoperability at present is that of the Semantic Web. This phrase, coined by Berners-Lee et al. (2001), envisages all resources on the Internet being joined by semantic linkages, which convey some meaning about their relationships with each other. The method to achieve this is to break down these relationships into small, atomistic components known as 'triples': these are so called, because they encode a 'subject', an 'object' and the semantic relationship between them (known as a 'predicate'). To record, for instance, that an individual donated an item to an archive, for instance, a triple of this type could be used (Table 1):

Table 1

Sample RDF triple to record donor of item

Subject	Predicate	Object
Michael Foot	Donated	Item 00250

Such a triple is relatively imprecise, however: which 'Michael Foot' is the subject, for instance? To obviate this imprecision, each component is usually given a unique identifier, known as a universal resource identifier (URI), which is defined in a controlled vocabulary or similar resource. Replacing the components of this triple with URIs produces this (Table 2):

Table 2

The same RDF triple as in Table 1 encoded with URIs

Subject	Predicate	Object
http://id.loc.gov/authorities/names/n50060550	http://cendari.eu/id/archive-actions/donation	http://thisarchive/items/00250

In this way, a complex web of semantic links can be built up, which can describe any resource and its relationship with any other. It is theoretically possible, for instance, to express all of the components and relationships within an EAD description in RDF [as has been attempted by, for instance, by a recent project in the UK (SALDA project n.d.)]. There are several reasons, however, for doubting the efficacy of RDF as a core format for archival description: these include the large number of triples required to encode even a simple description; potential intellectual property rights issues that may arise from a loss of control over data in RDF environments (Hawtin et al. 2011, p. 21); and curation problems arising from an incompatibility between the fluid boundaries of RDF data and the discrete packaging required by core digital preservation models such as open archival information system (OAIS) (Consultative Committee for Space Data Systems 2009).

If RDF triples do not form an ideal medium for storing archival descriptions, it would undoubtedly be useful for these to incorporate some of the features of RDF in order to enable wider interoperability of the type described here. Employing URIs to define concepts and their relationships, for instance, is a mechanism to allow semantic interoperability across the entire Internet, and such identifiers should form the primary mechanism for defining these precisely. Deconstructing concepts into more granular components, while preserving their relationships, allows interoperable connections to be made more precisely.

Reconciling the requirements of text with the more atomistic requirements of interoperability is perhaps the greatest challenge to achieve this. It is important not to lose the communicative power of the text description and reduce all components to atomistic data elements: many of the nuances of collection description,

particularly those relating to its scope and content, are not readily reduced in this way. A viable collection description architecture must retain both facets to capture the richness of its content. EAD incorporates both elements within its architecture by combining its descriptive components with an array of more constrained elements designed to emulate the fields of a traditional database structure: these include most elements found within its <did> (descriptive identification) section and the array of 33 sub-elements available to mark up the content of <p> (paragraph) elements.

These more data-like elements have proved useful in allowing the extraction and consolidation of metadata from multiple EAD files, and the creation of such union catalogues as AIM25 and the Archives Hub. At present, however, these union catalogues of archives offer only relatively basic searching. AIM25, for instance, allows searching by keyword, with limits by institution and date, in its advanced search options. The Archives Hub offers a richer range of search options, including keywords, creators, dates, subjects, names and media types, although these remain a smaller set than can be offered by systems based on more data-like metadata schemes (including union library catalogues based on MARC).

Much of the problem in providing catalogues of this type is the fact that so much information is embedded in the textual content of descriptions in finding aids and so is not addressable as data components. More importantly, in the context of the contemporary interoperable information environment, particularly that of the Semantic Web, EAD elements can prove problematic as they incorporate no unambiguously defined method for recording URIs. It may be possible to use these identifiers in some parts of the current EAD architecture (for instance, in the frequently recurring TYPE attribute), but this requires extending the intended semantics of most components.

These issues are being addressed by the EAD community, particularly in the form of current revisions being made to the standard under the banner EAD3 (Society of American Archivists 2014b). One proposal under consideration is the creation of 'EAD-Strict', a more constrained version of the EAD base schema, which will be more 'data-centric' and so more interoperable (Society of American Archivists 2014c, p. 3). Further proposals would increase the granularity of such components as dates and physical description (Society of American Archivists 2014a, pp. 32–35) and create a generic <relation> element to link the materials being described to another entity (Society of American Archivists 2014d, p. 111). These revisions will undoubtedly improve the interoperability of EAD documents (particularly EAD-Strict ones), although as presently proposed, they fall a little short of allowing archival descriptions to achieve the full interoperability of the Semantic Web or to become datasets capable of analysis rather than description. For this reason, the CENDARI project attempted to draw up the database-like schema described below.

The CENDARI project

The CENDARI project aims to provide more than a union catalogue of archival holdings and so requires a higher degree of integration of the components of archival descriptions than is necessitated by catalogues such as AIM25 or the Archives Hub. In particular, it aims to produce a 'digital ecosystem' for historical research, an enquiry environment in which metadata and resources can be linked together to produce a shared

virtual research infrastructure; within this techniques such as data mining, visualisation and online annotation can be employed co-operatively to produce dynamic networks of information, which change as research proceeds (Gartner and Hedges 2012, p. 1).

An important feature of the project is that archival descriptions should act as datasets, which can be subject to detailed analysis. To do so requires the seamless integration of components in archival descriptions with user-generated content (such as annotations) at fine levels of granularity; it also requires these descriptions not only to act as aids to discovery, but also to be capable of supporting machine-based techniques used more commonly for analysing large datasets or text corpora.

The CENDARI collection schema (CCS)

For many of the reasons outlined above, it was concluded that EAD alone, as it currently stands, would not be able to support these functions. It was decided therefore to design a new schema, which could extend its functionalities by encoding collection descriptions in a form more readily amenable to the dynamic analytical environment envisaged by the CENDARI project while still retaining EAD's descriptive functionality. The schema developed, the CENDARI collection schema (CCS), is designed to function independently, but also to act an 'intermediary schema' (Gartner 2011) to EAD. An intermediary schema is designed to act as a mediator to other schemas as much as a standalone standard. As such, CCS provides a more precise and constrained element set from which (relatively simple) EAD records can be generated in order to allow integration with existing systems. Potential ways in which EAD and CCS can co-operate are discussed later in this article.

The schema was constructed initially by a process of facet analysis. A number of domain practitioners were asked to identify key components or facets of an ideal collection-level record, divorced if possible from the element set provided by EAD. This exercise produced a set of top-level and second-level facets as follows (Fig. 1):

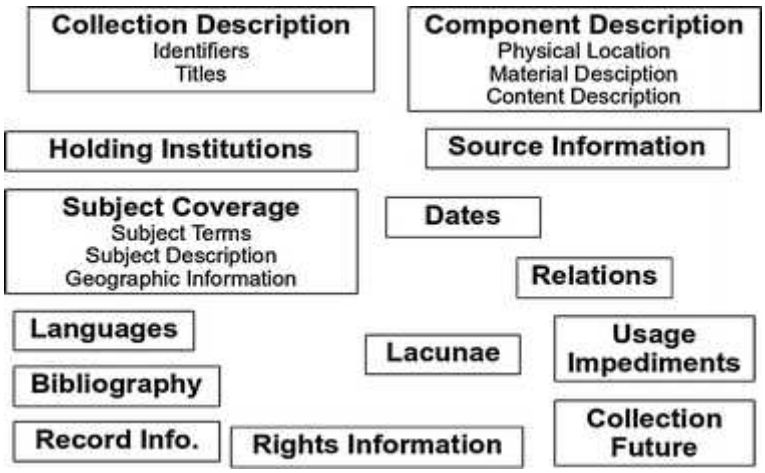


Fig. 1 CCS facets derived from initial facet analysis exercise

The semantic breadth of these top-level facets usually requires circumscribing to enable them to function as viable components of a description: <lacunae>, for instance, requires defining more precisely by type of lacuna to provide useful information to the custodians or users of a collection. This semantic narrowing is enabled for most components in the structure by the use of two attributes, *type* and *typeURI,* to qualify their respective elements:

```
<lacunae>
    <lacuna
      type="missing component"
       typeURI="http://cendari.edu/id/lacunatypes/missingcomponent">
    </lacuna>
</lacunae>
```

The *type* attribute is used throughout the schema to provide a human-readable term for the narrower semantic component of the broader facet, and *typeURI* to contain a URI for the *type;* both of these should preferably be taken from an ontology or controlled vocabulary.

Using this approach combines the advantages of a closely structured XML architecture with the semantic flexibility of an ontology. Interoperability, however, requires the use of widely used (preferably community-defined) ontologies, in which the URIs used are defined. At present, most components of the CENDARI description lack ontologies of this type; consequently, a major part of the project is the definition of a number of these to underpin its semantic architectures.

In addition to allowing the semantic refinement of core facets in the schema, the same methodology is used throughout to supply regularised forms of components: subject terms, for instance, make use of *reg* and *regURI* attributes to supply human-readable and URI-encoded versions of their content:

```
<subjectTerm regURI="http://id.loc.gov/authorities/subjects/sh85085001"
                reg="Middle Ages"/>
```

This approach is also used to encode essential sub-facets of each component. In the case of <lacuna>, for instance, the cause of any gap in a collection can be described in a *cause* attribute and its *causeURI* counterpart:

```
<lacuna
        cause="fire"
        causeURI="http://cendari/edu/id/lacunacauses/fire">
        <p>A major part of the collection was destroyed by fire</p>
</lacuna>
```

Every sub-facet may be defined using URIs in this way: these definitions may be extended to very fine levels of granularity, for instance in defining the units in which measurements are described:

```
<materialType
            unitSize="250x170"
            unitSizeUnits="mm"
            unitSizeUnitURI="http://cendari.eu/id/extentunit/millimetre"
             materialType="paper"
             materialTypeURI="http://cendari.eu/id/unittype/paper"
             extent="245"
             extentUnits="m"
             extentUnitURI="http://cendari.eu/id/extentunit/metre"
             unitCount="3795"
             unitCountUnits="leaf"
             unitCountUnitsURI="http://cendari.eu/id/unitcountunit/leaf"
             lang="en">
   <desc>
      <p>The bulk of the collection comprises 245 metres of boxed individual paper leaves</p>
   </desc>
</materialType>
```

As these examples show, textual descriptions of each component (here included in <p> (paragraph) elements) are also available for most elements.

A further important component of many elements in this schema is a facility to impose chronological limits on their validity or application. A lacuna in a collection, for instance, may have its scope marked by *startDate* and *endDate* attributes:

```
<lacuna
         startDate="1923-02-02"
         endDate="1924-12-12"
         calendar="gregorian">
       <p>Years 1923-24 are missing as a result of water damage</p>
    </lacuna>
```

All date components may be referenced in any calendar, although the Gregorian is the default option for these. This facility allows materials dated in, for instance, the Julian calendar, to be co-analysable with those in the Gregorian by the application of simple conversion algorithms.

The core of the CENDARI collection-level description is the component description, an analogue to the EAD container list. In a similar way to its EAD counterpart, this facet is arranged hierarchically in the shape of a top-level <collection> element containing up to six levels of nested <component> elements. The <collection> and <component> elements may all contain information on the physical location of their constituent components, a description of their material make-up (including a condition report), and a description of their contents (akin to the scope and content descriptions of EAD).

Each <collection> and <component> element is identified by an obligatory XML ID attribute by which it is referenced from other parts of the record. Most top-level facets include an attribute, *coverageID,* which records the parts of the collection (or the collection as a whole) to which they apply. A geographical term, for instance, may be limited to a part of a collection in this way:

```
<geogTerm reg="Germany"
        scheme="MARC countries"
        schemeURI="http://id.loc.gov/vocabulary/countries"
        regURI="http://id.loc.gov/vocabulary/countries/gw"
       coverageID="cendari-sample-1-component.1.2">Germany</geogTerm>
```

where its *coverageID* attribute contains the value of the *ID* of the component whose contents relate to the country cited.

This referencing methodology allows the coverage of every facet of the collection-level record to be delineated at a very fine level of granularity, so providing a highly detailed, machine-readable description of the collection.

The use of the *coverageID* attribute throughout the schema records interconnections between components, but to establish links outside the architecture of the single record, a facility exists to encode external relationships. The <relation> element offers the following structure:

```
<relation type="institution"
    typeURI="http:/cendari.eu/id/institution"
    target="European Imaginary Archive"
    targetURI="http:/cendari.eu/id/institution/281829"
    coverageID="cendari-sample-1-collection"/>
```

As before, URIs are used to identify the type of relationship and its target or object; *coverageID*is also used to limit the scope of the relationship to a part of the collection if necessary.

Enhancing access to collection descriptions using the CENDARI schema

From the above paragraphs, it is clear that the CENDARI schema attempts to extend the capabilities of EAD by introducing a set of discrete, semantically focused components. Every component of the schema supplements its descriptive (usually textual) content with an extensive set of attributes to allow the encoding of its core semantic sub-units at fine levels of granularity. Each of these attributes includes the facility to record URIs, so allowing their integration with external ontologies and from there to the wider Semantic Web environment.

The fine-levelled granularity of the CENDARI schema allows searches to be focussed on a wide set of potential elements. The total number of these is unlimited as many can be refined beyond their initial definition using *type* and *typeURI* attributes. Although similar qualification by type is possible for some EAD

elements (for instance <unittitle>), it is not at present as consistently applied as in the CENDARI schema and provides no facility to use URIs for precise semantic referencing.

This new set of semantically focused elements and attributes should allow the creation of more detailed and sophisticated search interfaces than currently exist, perhaps the most obvious benefit that could accrue from the features of CCS. Beyond its use as an enhanced discovery tool, however, the CENDARI schema should move towards fulfilling the requirements of the CENDARI project for archival descriptions to act as datasets to support machine-readable analysis. At the time of writing, research is only just beginning with the project team to assess the types of analysis that may be possible (and how useful they could be for archivists and researchers). But, to provide an example of one potential type, it would be possible (for conservationists, for instance) to use the <lacuna> element described earlier to produce detailed statistical analyses of patterns of damage to collection holdings. A complete <lacuna> element of this type allows a number of analyses to be carried out:

```
<lacuna
        type="missing component"
        typeURI="http://cendari.edu/id/lacunatypes/missingcomponent"
        cause="fire"
        causeURI="http://cendari/edu/id/lacunacauses/fire"
        coverageID="cendari-sample-1-component1"
        startDate="1923-02-02"
        endDate="1924-12-12"
        calendar="gregorian">
        <p>A major part of the collection was destroyed by fire</p>
</lacuna>
```

It is immediately possible to use metadata encoded in this way to analyse the types and causes of gaps in collections and to form a diachronic picture of them. Because this element is linked by its *coverageID* to the components of the collection to which it applies, it is possible to derive information on the types of materials involved, their physical location, their provenance and the actions performed on them. This information can then be readily correlated and analysed for patterns. When a significant set of records are available in this schema, the potential for extensive analyses across archives becomes particularly valuable. Little empirical data of this type is currently available in a form, which makes itself amenable to large-scale computational analysis.

A further potential use for these collection descriptions is as datasets to support dynamic research and resource guides. As part of the CENDARI project, a number of scholars have been examining how its research infrastructure can be used for their specific areas of historical research. One has proposed a structure for online research guides, which can be dynamically constructed from a bank of collection descriptions (Salvador 2013). The main components suggested for such a guide are as follows (Fig. 2):

Fig. 2

Suggested components of an online research guide (from Salvador 2013, slide 11)

All of these components can readily be generated from the CENDARI schema by searching and extracting metadata at an appropriate level of granularity. **Time span** components, for instance, may be generated using the *startDate* and *endDate* attributes available for all elements with a potential temporal component. **Types of sources** are readily extracted from the schema's material description elements, which can use ontologies to provide normalised source types. **Links to online search engines** are readily encoded using the generic <relation> element described above.

When multiple archives are described with this schema, the online research guide can become a potentially substantial cross-archival resource updated dynamically as these collections develop.

It should be emphasised that, at the time of writing, work is only beginning within the CENDARI project to examine potential applications of the CCS schema, and the possibilities outlined above will be supplemented by later work, which will probe these questions in more detail.

Integration with the semantic web

Although the CENDARI schema uses the hierarchical architecture of XML for many of the reasons outlined earlier, its consistent use of URIs for semantic definition makes it readily compatible with the methodologies of the Semantic Web (Berners-Lee et al. 2001). It may be useful to generate triples from the CENDARI XML schema to allow enhanced access to collection descriptions as linked open data. By outputting these to a 'triple store' and allowing access to this as an endpoint for the RDF query language SPARQL Protocol and RDF Query Language (SPARQL) (W3C Consortium 2008), it is possible for users to interrogate the record in ways, which need not be predicted in advance by the designers of a query environment.

Generating triples from the schema is achieved relatively simply using standard XSLT transformations. The simple <subjectTerm> element noted earlier, for instance:

```
<subjectTerm regURI="http://id.loc.gov/authorities/subjects/sh85085001"
             reg="Middle Ages"/>
```

may be converted into triples of this form (Table 3):

Table 3

<subjectTerm> attributes rendered as RDF

Subject	Predicate	Object
http://cendari.eu/id/collection/sample-1/subjectTerm/00001	http://cendari.eu/id/predicates/hasRegURI	http://id.loc.gov/authorities/subjects/sh85085001
http://cendari.eu/id/collection/sample-1/subjectTerm/00001	http://cendari.eu/id/predicates/hasReg	'Middle Ages'

The entire collection description can readily be output as a set of triples without manual input. Ontologies may then be used to link the CENDARI-specific URIs (particularly those used for predicates) to more generic pre-existing ontologies and from there to the Semantic Web generally.

Compatibility with existing archival standards

Although designed with the specific purpose of supporting the CENDARI research infrastructure, compatibility of the schema with existing archival practice is important for both the project and its potential use beyond it. Specifically, it should conform if at all possible to the requirements of ISAD(G) and be able to interoperate as far as possible with EAD if it is to be of use to the archives community.

A mapping of the requirements of ISAD(G) with the CCS schema reveals a high level of conformance. The four rules of multi-level description prescribed by ISAD(G) [description from the general to the specific, information relevant to the level of description, linking of descriptions and the non-repetition of information (International Council on Archives 2000, p. 12)] are readily accommodated. The first of these, which mandates the overall structure of a description, must be applied by aggregating components distributed throughout the CCS file; the others are all supported by its architecture.

The 26 elements of description specified by ISAD(G) (International Council on Archives 2000, pp. 15–35) in general map well to the CCS schema (Table 4):

Table 4

Mapping of ISAD(G) 26 elements of description to CCS

ISAD(G) element of description	**CCS mapping**
Identity statement area	
Reference code	<collectionIdentifier> and <physicalLocation> sub-element
Title	<title> sub-elements of <collection> or component elements
Dates	<date> element
Level of description	*level* attribute of <collection> or component elements
Extent and medium of the unit of description	sub-elements of <materialDesc> and <materialType>
Context area	
Name of Creator	<provenance> element with *role* attribute set to 'collection creator'
Administrative/biographical history	<provenance> element allows text descriptions (including biographical information), although this may be better recorded in external schemas referenced by *regURI* attribute
Archival history	<provenance> elements (repeatable)
Immediate source of acquisition or transfer	<provenance> elements (repeatable)
Content and structure area	
Scope and content	<contentDesc> sub-element of <collection> or component elements
Appraisal, destruction and scheduling information	<collectionFuture> (only basic information at present)
Accruals	<collectionFuture>
System of Arrangement	Maybe be recorded in <contentInfo> or else expressed by structure of component elements
Conditions of access and use area	

Conditions governing access	<rightsInformation>
Conditions governing reproduction	<rightsInformation>
Language/scripts of material	<language>
Physical characteristics and technical requirements	<materialDesc> sub-element of <collection> or component elements
Finding aids	*source* and related attributes of <contentInfo>
Allied materials area	
Existence and location of originals	<relation> element with appropriate *typeURI* attribute
Existence and location of copies	<relation> element with appropriate *typeURI* attribute
Related units of description	<relation> element with appropriate *typeURI* attribute
Publication note	<bibliography>
Notes area	
Note	No specific element for information that cannot be covered by existing elements and attributes: recommend use of <relation> element for this
Description control area	
Archivist's Note	<note> sub-elements within <recordInformation>
Rules or conventions	<note> sub-elements within <recordInformation>
Date of description	<creationDate> and <changeDate> within <recordInformation>

The most notable broken mapping is the generic *note* element in ISAD(G), which provides a container for information not readily accommodated in more specific elements. At present, no such element is present within CCS. Instead, it is recommended to use a <relation> element to reference an external note, although a future revision of the schema could readily accommodate this. Other mappings rely on the use of suitable *typeURI* attributes to provide a match to ISAD(G): **archivist's note** and **rules or conventions**, for instance, use the same <note> elements, which need to be qualified in this way to differentiate them.

While some further development of the schema may therefore be necessary to ensure full ISAD(G) compliance, this should be relatively minor. A dialogue with the archival community should underlie this development to ensure full compatibility.

Crosswalking to the more flexible requirements of EAD is more problematic. As part of the project, a detailed mapping to EAD was carried out and an XSLT stylesheet designed to translate instances of this schema to EAD. This mapping is relatively easy to achieve in the case of those components, which map to the more data-like elements of EAD (including such elements as <unittitle>, <unitdate> or <unitid>). For instance, the following collection-level information in the schema translates to components of the <did> within the <archdesc> of an EAD finding aid (Fig. 3):

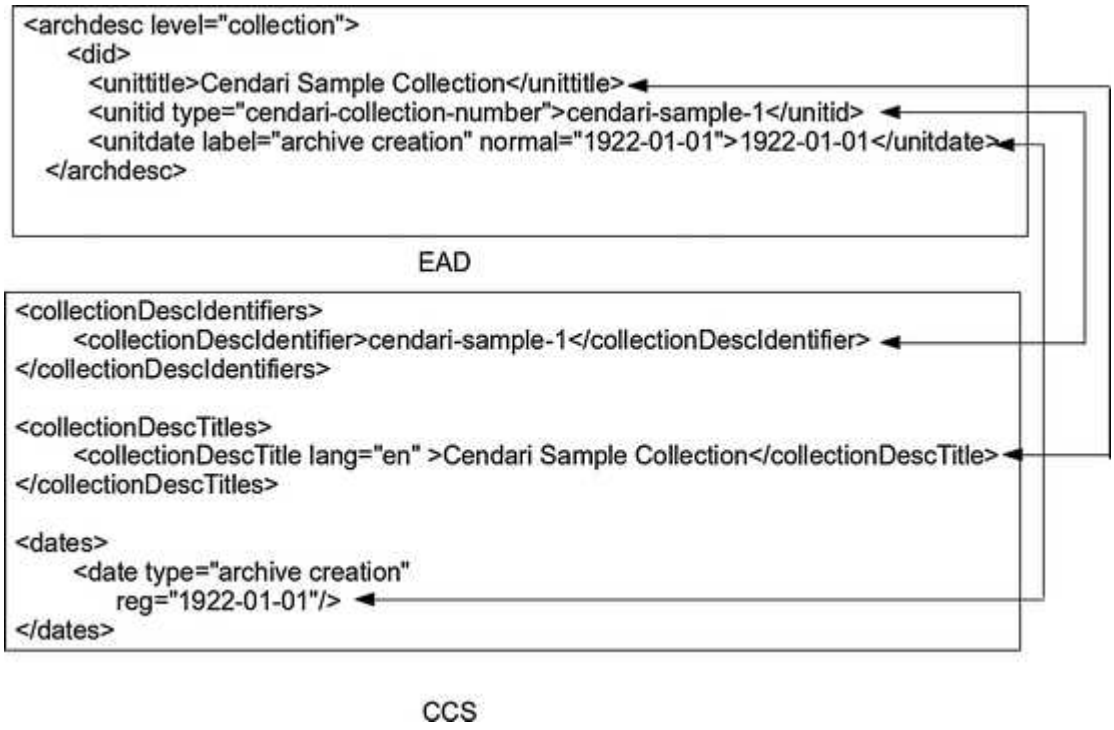


Fig. 3

Mapping of collection-level information between EAD and CCS

However, in the case of many EAD elements such as <language>, the CENDARI schema offers a more extensive attribute set, which does not translate fully to its counterpart. A typical CENDARI <language> element may take this form:

```
<language
        ID="ccs-instance-element.123"
        coverageID="cendari-sample-1-component1"
        schemeURI="http://id.loc.gov/vocabulary/iso639-1"
        scheme="ISO-639"
        reg="de"
        regURI="http://id.loc.gov/vocabulary/iso639-1/de"
        sourcePerCent="25"
        script="Latn"
        scriptURI="http://lexvo.org/id/script/Latn"
        scriptScheme="ISO 15924:2004"
        scriptSchemeURI="http://standards.iso.org/iso/15294-2004">German</language>
```

This example includes such information as the scheme from which the language code (given in *reg*) is taken, the percentage of the materials in the language, the script used, the scheme from which the script name is taken and URIs for each of these concepts. By contrast the EAD <language> element only offers *langcode* (language code) and *scriptcode* (script code) attributes to cover these concepts.

Crosswalking from EAD to CCS is likely to experience similar problems. The more unitary EAD elements, such as <unitttitle>, <unitdate> and <unitid> discussed above, can readily populate their more precise counterparts in CCS. Because, however, so much of an EAD record is descriptive prose, most of it can only be included as textual element content within CCS which effectively ignores its facility for more precise semantic referencing offered by the attribute sets detailed above. Nonetheless, a number of possible methods can be used to enable CCS to extend or complement EAD.

Working with EAD

One possible means of using CCS in conjunction with EAD is to employ it as an 'intermediary schema'. An intermediary schema is defined in Gartner (2011) as a heavily constrained XML schema designed to facilitate the creation of metadata for a specific application which can also be used to generate an instance of a more flexible and complex established schema using XSLT transformations. This is particularly useful where the target schema, such as EAD, offers multiple ways to encode similar information.

CCS can function as such an intermediary schema: a crosswalk has been written between CCS and EAD. As mentioned earlier, an XSLT stylesheet has already been written to generate EAD instances from CCS files. For the reasons outlined above, this does, however, generate a relatively constrained EAD instance, omitting several frequently used elements (particularly those supporting the descriptive components outside the container listing). This limited EAD instance may, however, fulfil the functions envisaged for the limited EAD-Strict element set under consideration by the EAD3 revision team. If so, it could prove a useful method for generating EAD-Strict instances.

An alternative approach would be to use CCS as a method for enriching or extending EAD by using the added features offered by its element and attribute sets. A number of possible methods could be used to do this. One may be to embed CCS data directly within an EAD file, although this is problematic as EAD does not allow any extensions [similar to the <extension> element available within the MODS (Metadata Object Description Schema) schema (Library of Congress 2010)] to enable XML data to be embedded within it. One potential means to overcome this is to escape components of XML syntax (such as the <and> symbols) and embed the resulting text within a <p> element. This could be done, for instance, using the frequently occurring <odd> element as follows:

```
<odd type="usage impediment - CCS">
<p>&lt;usageImpediment
        typeURI="http://cendari.eu/id/impediment/illegibility"
        type="illegibility"
        lang="en"
        severity="75" &gt;Approx. 75% of texts illegibility owing to mice damage.&lt;/usageImpediment&gt;
</p>
</odd>
```

An alternative approach would be to convert a CCS instance to RDF as discussed above and enrich the EAD file with the generated triples: these could, for instance, be encoded as a <list> within <odd>:

```
<c01>
<odd type="subjectTerm-rdf-statements">
    <list>
<item>&lt;http://cendari.eu/id/collection/sample-1/subjectTerm/00001&gt;
        &lt;http://cendari.eu/id/predicates/hasRegURI&gt;
        &lt;http://id.loc.gov/authorities/subjects/sh85085001&gt; .</item>

<item>&lt;http://cendari.eu/id/collection/sample-1/subjectTerm/00001&gt;
        &lt;http://cendari.eu/id/predicates/hasReg&gt;
        'Middle Ages' .</item>
    </list>
</odd>
</co1>
```

A final approach may be to use the CCS file as a parallel document and use EAD's linking facilities to reference components within it. For instance, detailed information on a language used in a collection, encoded in CCS, could be referenced from an EAD <language> element as follows (Fig. 4):

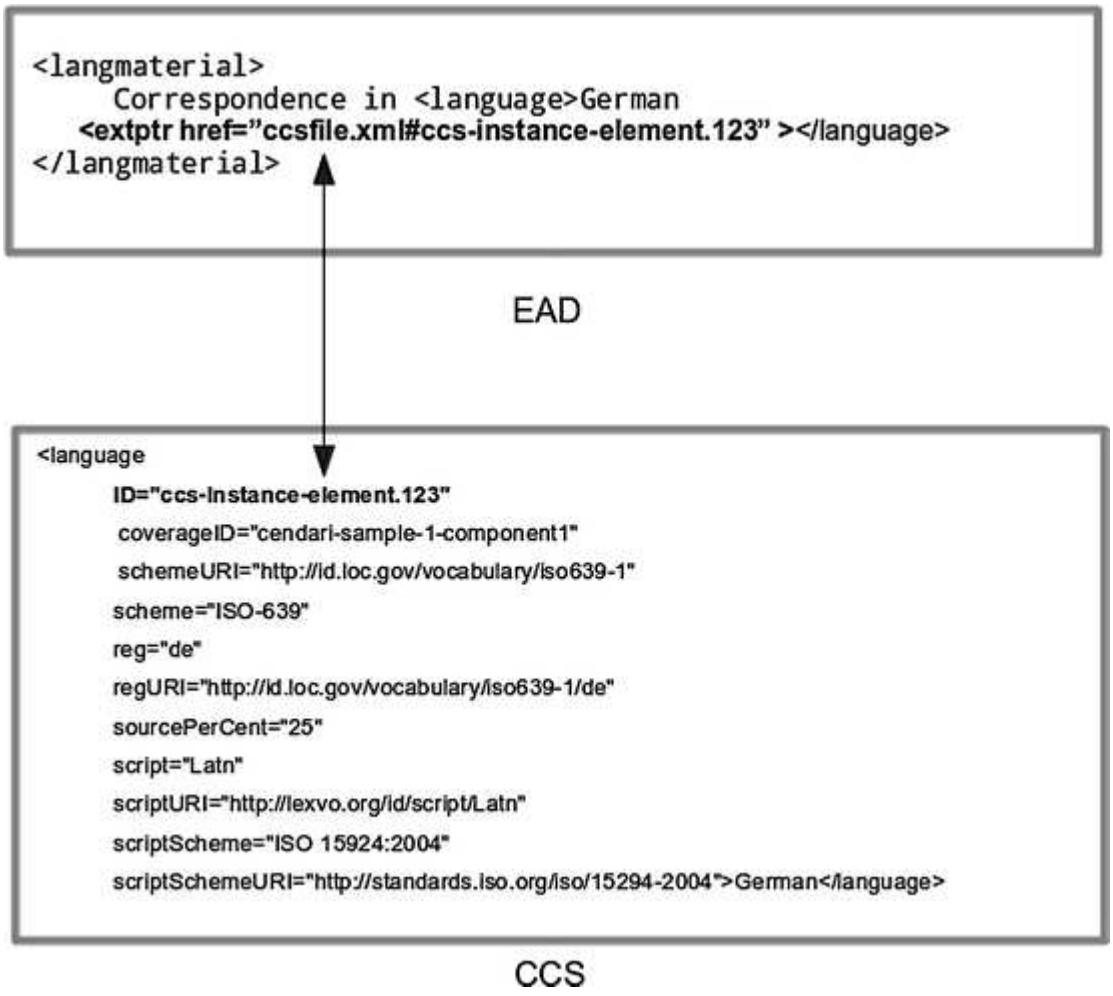


Fig. 4

Using EAD's <extptr> to link to a CCS element

None of these approaches is ideal, but all offer possibilities for linking EAD and CCS in the forms in which both standards currently exist. It is, of course, possible that future revisions of EAD may incorporate some of the features advocated here to enhance its interoperability: the consistent application of URIs (akin to the *regURI* and *typeURI* attributes in CCS) may well enhance its semantic foundations and so make it more compatible with the Semantic Web. Any such future changes will, of course, depend on how the archival community decides to approach the challenges laid down by linked data, to which discussion of this article will hopefully contribute.

Conclusions

The schema outlined in this article appears a viable approach towards enhancing the descriptive functionality of EAD to incorporate more semantically interoperable components; this could potentially allow many of the advantages of Semantic Web technologies to become applied to archival descriptions. In doing so, it goes some way to meeting aspirations, such as Pitti's discussed earlier, to combine database and markup functions within archival descriptions, aspirations which are currently under community-based discussion in the form of EAD revisions.

Incorporating these features into an archival description offers the potential for enhanced searching and browsing facilities at finer levels of granularity than has often been possible before. More radically, it offers the potential for archival descriptions to become more tractable resources capable of machine-readable analysis on large corpora of metadata. A large body of records encoded in this schema offers the potential to achieve more comprehensive and precise empirical data on archival holdings. This could potentially benefit the custodians of collections as much as the scholars who rely on them for their research.

This research is still relatively new, and at the time of writing, evaluative testing of the standard in the context of the CENDARI architecture and the workflows supporting it is only in its early stages. It can therefore only be offered as a heuristic proposal at present and will benefit greatly from feedback from the archival and data management communities. As these evaluation and feedback are received, the schema will undoubtedly be revised, but it is hoped that it has raised important issues, which have a bearing on the future direction of archival description and has offered a putative way of taking these forward.

Notes

Acknowledgments

The research leading to this article has received funding from the European Union's Seventh Framework Programme [FP7/2007-2013] under Grant Agreement No 284432. The author gratefully acknowledges the contributions made to this project by the other team members, in particular David Stuart and Sheila Anderson from King's College London.

References

Berners-Lee T et al (2001) The semantic web. Sci Am 284(5):28–37

Bunn J (2013) Developing descriptive standards: a renewed call to action. J Soc Arch 34(2):235–247

Cendari Project (2013) Cendari | Collaborative European Digital Archive Infrastructure http://www.cendari.eu/. Accessed 19 February 2013

Consultative Committee for Space Data Systems (2009) Reference Model for an Open Archival Information System (OAIS): draft recommended standard. Accessed 19 April 2011 at: http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf

Dow EH (2009) Encoded Archival Description as a halfway technology. J Arch Organ 7(3):108–115

Gartner R (2011) Intermediary schemas for complex XML publications: an example from research information management. J Digit Inf 12. http://journals.tdl.org/jodi/article/view/2069. Accessed 07 July 2012

Gartner R, Hedges M (2012) CENDARI: establishing a digital ecosystem for historical research. In: 7th IEEE international conference on digital ecosystems and technologies proceedings

Hawtin R et al. (2011) Review of the evidence for the value of the 'linked data' approach: final report to JISC. Accessed 27 July 2012 at: http://ie-repository.jisc.ac.uk/559/1/JISC_Linked_Data_Review_Oct2011.pdf

International Council on Archives (2000) ISAD(G): General international standard archival description, Second edition. Accessed 26 April 2014 at: http://www.ica.org/en/node/30000

King's College London Archives (n.d.) AIM25—online research for archive collections of higher education institutions and learned societies within greater London. Accessed 01 November 2013 at: http://www.aim25.ac.uk/

Library of Congress (2010) MODS (Metadata Object Description Schema) Official Web Site http://www.loc.gov/standards/mods. Accessed 29 April 2014

Library of Congress (2011) EAD: Encoded Archival Description Version 2002 Official Site (EAD Official Site, Library of Congress). http://www.loc.gov/ead/. Accessed 29 November 2011

MIMAS (n.d.) Home—Archives Hub. http://archiveshub.ac.uk/. Accessed 1 November 2013

Pitti D (1999) Encoded Archival Description: an introduction and overview. D-Lib Mag 5/11. http://www.dlib.org/dlib/november99/11pitti.html. Accessed 28 January 2010

SALDA project (n.d.) SALDA project: Sussex archive linked data application. Accessed 17 April 2014 at: http://blogs.sussex.ac.uk/salda/2011/07/25/converting-ead-data-to-rdf-linked-data/

Salvador A (2013) They're reading our minds: humanities research and digital thinking with CENDARI. Accessed 2 December 2013 at: http://www.youtube.com/watch?v=Dj0rmefDzNM

Shaw EJ (2009) Rethinking EAD: balancing flexibility and interoperability. New Rev Inf Netw 71:117–131

Society of American Archivists (2014a) EAD revision progress report. www2.archivists.org/sites/all/files/eadRevisionProgress_2013-08-16.pptx. Accessed 23 April 2014

Society of American Archivists (2014b) EAD Revision. http://www2.archivists.org/groups/technical-subcommittee-on-encoded-archival-description-ead/ead-revision. Accessed 22 April 2014

Society of American Archivists (2014c) EAD—Technical Considerations. http://www2.archivists.org/sites/all/files/EADRevisionTechnicalConsiderations_0.pdf. Accessed 22 April 2014

Society of American Archivists (2014d) EAD Revised Tag Library. http://www2.archivists.org/sites/all/files/Elements_Beta.pdf. Accessed 22 April 2014

W3C Consortium (2008) SPARQL query language for RDF. Accessed 10 June 2013 at: http://www.w3.org/TR/rdf-sparql-query/

Bibliography

This bibliography lists all items cited within this critical summary: it supplements references cited within the submitted articles which are not duplicated here.

Abiteboul, S. et al. (2001) 'Xyleme, a dynamic warehouse for XML data of the Web', in *Database Engineering and Applications, 2001 International Symposium on.* 2001 IEEE. pp. 3–8.

Abiteboul, S. et al. (2002) The Xyleme project. *Computer Networks*. 39 (3), 225–238.

Allinson, J. et al. (2007) A Dublin Core application profile for scholarly works. *Ariadne*. (50) [online]. Available from: http://www.ariadne.ac.uk/issue50/allinson-et-al (Accessed 24 April 2017).

Almarimi, A. & Pokorny, J. (2005) A mediation layer for heterogeneous XML schemas. *International Journal of Web Information Systems*. 1 (1), 25–33.

Anderson, S. (2013) What are research infrastructures? *International Journal of Humanities and Arts Computing*. 7 (1–2), 4–23.

Anderson, S. & Blanke, T. (2012) Taking the long view: from e-science humanities to humanities digital ecosystems. *Historical Social Research/Historische Sozialforschung*. 3, 147–164.

Archivio Centrale dello Stato (2017a) *OAD Vocabulary Specification 1.2* [online]. Available from: http://labs.regesta.com/progettoReload/wp-content/uploads/2013/08/oadNew.html#OAD (Accessed 20 June 2017).

Archivio Centrale dello Stato (2017b) *Reload: Obiettivi* [online]. Available from: http://labs.regesta.com/progettoReload/en/presentazione/ (Accessed 20 June 2017).

Barbosa, D. et al. (2001) 'ToX-the Toronto XML Engine.', in *Workshop on Information Integration on the Web*. 2001 pp. 66–73. [online]. Available from: http://www.pms.ifi.lmu.de/publikationen/projektarbeiten/Felix.Weigel/xmlindex/material/barbosa01tox.pdf (Accessed 5 March 2017).

Barbosa, D. et al. (2002) 'ToXgene: a template-based data generator for XML', in *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*. 2002 ACM. pp. 616–616. [online]. Available from: http://www.cs.toronto.edu/tox/toxgene/docs/ToXgene.pdf (Accessed 5 March 2017).

Bauman, S. (2011) 'Interchange vs interoperability', in *Balisage: The Markup Conference 2011: Proceedings*. 2011 [online]. Available from: https://www.balisage.net/Proceedings/vol7/html/Bauman01/BalisageVol7-Bauman01.html (Accessed 19 April 2017).

Beer, C. A. et al. (2009) 'Developing a flexible content model for media repositories: a case study', in *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*. 2009 ACM. pp. 97–100.

Blanke, T. et al. (2017) The European Holocaust Research Infrastructure Portal. *Journal on Computing and Cultural Heritage (JOCCH)*. 10 (1), 1–17.

Blanke, T. & Kristel, C. (2013) Integrating Holocaust research. *International Journal of Humanities and Arts Computing*. 7 (1–2), 41–57.

Bohring, H. et al. (2005) Mapping XML to OWL ontologies. *Leipziger Informatik-Tage*. 72, 147–156.

Boley, H. & Chang, E. (2007) 'Digital ecosystems: principles and semantics', in *DEST'07. Inaugural IEEE-IES*. 2007 pp. 398–403.

Bountouri, L. & Gergatsoulis, M. (2009) Interoperability between archival and bibliographic metadata: An EAD to MODS crosswalk. *Journal of Library Metadata*. 9 (1–2), 98–133.

Briscoe, G. & De Wilde, P. (2006) 'Digital ecosystems: evolving service-orientated architectures', in *Proceedings of the 1st international conference on Bio inspired models of network, information and computing systems*. 2006 ACM. pp. 17–23.

Buckland, M. K. (1997) What is a "document"? *Journal of the American Society for Information Science*. 48 (9), 804–809.

Buckland, M. K. (1998) What is a "digital document"? *Document Numérique*. 2 (2), 221–230.

Burnard, L. & Sperberg-McQueen, C. M. (1995) *TEI Lite: an introduction to text encoding for interchange*. Utrecht: SURFnet.

Burnard, L. & Sperberg-McQueen, C. M. (2006) *TEI Lite: encoding for interchange: an introduction to the TEI. Revised for TEI P5 release* [online]. Available from: http://www.tei-c.org/Vault/P5/1.5.0/doc/tei-p5-exemplars/pdf/teilite.doc.pdf (Accessed 16 August 2017).

Carey, M. J. et al. (2000a) 'XPERANTO: A middleware for publishing object-relational data as XML documents', in *Proceedings of the 26th International Conference on Very Large Databases*. 2000 VLDB. pp. 646–648. [online]. Available from: ftp://ftp.cse.buffalo.edu/users/azhang/disc/disc01/cd1/out/papers/vldb/xperantomiddlewmijej.pdf.

Carey, M. J. et al. (2000b) 'XPERANTO: publishing object-relational data as XML', in *WebDB (Informal Proceedings)*. 2000 pp. 105–110. [online]. Available from: http://db.ucsd.edu/cse232B-s05/papers/carey00xperanto.pdf (Accessed 5 March 2017).

CENDARI Project (2013a) *CENDARI | Collaborative European Digital Archive Infrastructure* [online]. Available from: http://www.cendari.eu/ (Accessed 19 February 2013).

CENDARI Project (2013b) *CENDARI: Guidelines for applying the schema* [online]. Available from: http://www.cendari.eu/sites/default/files/CENDARI_D6.2%20Guidelines%20for%20applying%20the%20schema.pdf (Accessed 4 May 2017).

Centre for e-Research (2011a) *BRIL* [online]. Available from: http://bril.cerch.kcl.ac.uk/ (Accessed 21 June 2011).

Centre for e-Research (2011b) *R4R: Readiness for REF* [online]. Available from: http://r4r.cerch.kcl.ac.uk/ (Accessed 7 February 2011).

Chan, L. M. (2005) 'Metadata interoperability: a methodological study', in *Proceedings of the Third China-US Library Conference, Shanghai, China*. 2005 pp. 23–25.

Chan, L. M. & Zeng, M. L. (2006) Metadata interoperability and standardization–a study of methodology part I. *D-Lib magazine*. 12 (6) [online]. Available from: http://www.dlib.org/dlib/june06/chan/06chan.html (Accessed 25 April 2017).

Cluet, S. et al. (2001) Views in a large scale XML repository. *The VLDB Journal—The International Journal on Very Large Data Bases*. 1, pp. 271–280.

Consultative Committee for Space Data Systems (2012) *Reference model for an Open Archival Information System (OAIS)* [online]. Available from: https://public.ccsds.org/pubs/650x0m2.pdf (Accessed 7 June 2017).

Cox, M. (2011) *Readiness for REF* [online]. Available from: http://www.rsp.ac.uk/documents/get-uploaded-file/?file=MarkCox_R4R.ppt (Accessed 22 May 2017).

DARIAH Project (2017) *DARIAH-EU* [online]. Available from: http://www.dariah.eu/ (Accessed 21 June 2017).

De Biagi, L. et al. (2012) Research product repositories: strategies for data and metadata quality control. *The Grey Journal*. 8 (2), 83–94.

Decker, S. et al. (2000) The semantic web: the roles of XML and RDF. *IEEE Internet Computing*. 4 (5), 63–73.

Deutsch, A. et al. (1998) *Xml-ql: a query language for XML* [online]. Available from: https://www.w3.org/TR/1998/NOTE-xml-ql-19980819/ (Accessed 17 August 2017).

DiLauro, T. et al. (2005) The archive ingest and handling test. *D-Lib Magazine*. 11 (12). [online]. Available from: http://webdoc.sub.gwdg.de/edoc/aw/d-lib/dlib/december05/choudhury/12choudhury.html (Accessed 19 April 2017).

Dow, E. H. (2009) Encoded Archival Description as a halfway technology. *Journal of Archival Organization*. 7 (3), 108–115.

Dublin Core Metadata Initiative (2004) *DC-Library Application Profile (DC-Lib)* [online]. Available from: http://dublincore.org/documents/library-application-profile/ (Accessed 27 August 2015).

Duval, E. et al. (2006) 'Application profiles for learning', in *Advanced Learning Technologies, 2006. Sixth International Conference on*. 2006 IEEE. pp. 242–246. [online]. Available from: https://lirias.kuleuven.be/bitstream/123456789/132968/1/425Duv.pdf (Accessed 26 April 2017).

EHRI Project (2017) *European Holocaust Research Infrastructure* [online]. Available from: https://www.ehri-project.eu/ (Accessed 21 June 2017).

EuroCRIS (2010) *CERIF Introduction* [online]. Available from: http://www.eurocris.org/Index.php?page=CERIFintroduction&t=1 (Accessed 8 February 2011).

EuroCRIS (2012a) *CERIF 1.3 | euroCRIS* [online]. Available from: http://www.eurocris.org/cerif-13 (Accessed 28 March 2017).

EuroCRIS (2012b) *CERIF 1.4 | euroCRIS* [online]. Available from: http://www.eurocris.org/cerif-14 (Accessed 28 March 2017).

Fan, C. et al. (2002) *XPERANTO: Bridging relational technology and XML*. San Jose: IBM Almadan Research Center.

Farquhar, A. & Hockx-Yu, H. (2008) Planets: integrated services for digital preservation. *International Journal of Digital Curation*. 2 (2), 88–99.

Fedora Commons (2002) *The Fedora Content Model Architecture (CMA)* [online]. Available from: http://fedora-commons.org/documentation/3.0b1/userdocs/digitalobjects/cmda.html (Accessed 9 December 2011).

Fedora Commons (2014) *Content Model Architecture - Fedora 3.8 Documentation - DuraSpace Wiki* [online]. Available from: https://wiki.duraspace.org/display/FEDORA38/Content+Model+Architecture (Accessed 6 April 2017).

Fedora Commons (n.d.) *The Fedora RELS-EXT ontology* [online]. Available from: http://www.fedora.info/definitions/1/0/fedora-relsext-ontology.rdfs (Accessed 1 June 2017).

Fernandez, M. F. et al. (2000) SilkRoute: trading between relations and XML. *Computer Networks*. 33 (1), 723–745.

Fernandez, M. F. et al. (2001) Publishing relational data in XML: the SilkRoute approach. *IEEE Data Engineering Bulletin* 24 (2), 12–19.

Gartner, R. (2016) *Metadata: shaping knowledge from antiquity to the Semantic Web*. Basel: Springer-Verlag.

Gartner, R. et al. (2013) A CERIF-based schema for encoding research impact. *The Electronic Library*. 31 (4), 465–482.

Gartner, R. & Grace, S. (2010) 'Modelling national research assessments in CERIF', in *Connecting science with society: the role of research information in a knowledge-based society: 10th International Conference on Current Research Information Systems*. 2010 pp. 97–105.

Gartner, R. & Hedges, M. (2013) 'CENDARI: establishing a digital ecosystem for historical research', in *7th IEEE International Conference on Digital Ecosystems and Technologies*. 2013 pp. 61–65.

Giaretta, D. (2008) The CASPAR approach to digital preservation. *International Journal of Digital Curation*. 2 (1), 112–121.

Godby, C. J. et al. (2003) 'Two paths to interoperable metadata', in *International Conference on Dublin Core and Metadata Applications*. pp. 19–27. [online]. Available from: http://dcpapers.dublincore.org/pubs/article/view/730.

Godby, C. J. et al. (2008) Toward element-level interoperability in bibliographic metadata. *Code4Lib Journal*. 1 (2) [online]. Available from: http://journal.code4lib.org/articles/54.

Goodkin, J. (2004) *OCLC's Digital Archive – disseminating with METS* [online]. Available from: https://www.loc.gov/standards/mets/presentations/od2/goodkin.ppt (Accessed 17 August 2017).

Gorman, P. & Prater, S. (2009) 'Beyond the tutorial: complex content models in Fedora 3', in *4th International Conference on Open Repositories*. 2009 Atlanta: Georgia Institute of Technology [online]. Available from: https://smartech.gatech.edu/bitstream/handle/1853/28479/145-644-1-PB.doc (Accessed 30 March 2017).

Gradmann, S. (2010) 'Digital library metadata', in *Business Planning for Digital Libraries: International Approaches*. Leuven: Leuven University Press. pp. 157–166.

Guenther, R. (2008) Battle of the buzzwords: flexibility vs. interoperability when implementing PREMIS in METS. *D-Lib Magazine*. 14 (7/8) [online]. Available from: http://www.dlib.org/dlib/july08/guenther/07guenther.html (Accessed 7 March 2014).

Guenther, R. & Myrick, L. (2006) Archiving web sites for preservation and access: MODS, METS and MINERVA. *Journal of Archival Organization*. 4 (1–2), 141–166.

Habing, T. & Cole, T. (2008) *ORE Resource Map implementation in METS: draft proposal* [online]. Available from: https://www.ideals.illinois.edu/bitstream/handle/2142/18685/ORE%201structMap_1REM.pdf (Accessed 16 March 2017).

Halevy, A. Y. et al. (2003) 'Piazza: data management infrastructure for semantic web applications', in *Proceedings of the 12th international conference on World Wide Web*. 2003 ACM. pp. 556–567.

Hawtin, R. et al. (2011) *Review of the evidence for the value of the 'linked data' approach: final report to JISC* [online]. Available from: http://ie-repository.jisc.ac.uk/559/1/JISC_Linked_Data_Review_Oct2011.pdf (Accessed 27 July 2012).

Heery, R. & Patel, M. (2000) Application profiles: mixing and matching metadata schemas. *Ariadne*. (25) [online]. Available from: http://www.ariadne.ac.uk/Issue25/App-Profiles/ (Accessed 17 August 2017).

Higher Education Funding Council for England (2007) *RAE2008 Data collection user guide* [online]. Available from: http://www.rae.ac.uk/pubs/2007/01/userguide/RAE%20User%20Guide.pdf (Accessed 28 September 2011).

Higher Education Funding Council for England et al. (2015) *Results & submissions : REF 2014 : outputs (REF2)* [online]. Available from: http://results.ref.ac.uk/Submissions/OutputsList/2198/Page1?searchId=447 (Accessed 2 August 2017).

Hitruhhina, J. (2012) *Digital ecosystem: a metaphor or the new type of ecosystem that uses ecological principles | [im]probability theory* [online]. Available from: https://imkerina.wordpress.com/2012/12/02/digital-ecosystem-a-metaphor-or-the-new-type-of-ecosystem-that-uses-ecological-principles/ (Accessed 13 June 2017).

Hochedlinger, N. et al. (2015) Standardized data sharing in a paediatric oncology research network—a proof-of-concept study. *Studies in Health Technology and Informatics*. 21, 227–34.

ICAR - Centro MAAS (2014) *SAN Ontologia* [online]. Available from: http://www.maas.ccr.it/SAN-LOD/lode/ (Accessed 20 June 2017).

Ingram, B. (2009) *ECHO Dep METS Profile for Master METS Documents* [online]. Available from: http://www.loc.gov/standards/mets/profiles/00000029.xml (Accessed 17 June 2011).

Jacinto, M. H. et al. (2004) XCSL: XML Constraint Specification Language. *CLEI Electronic Journal*. 6 (1: Paper 1), 1–29.

Kaiser, G. E. (2004) *Coping with complexity: A standards-based kinesthetic approach to monitoring non-standard component-based systems*. Defense Technical Information Center [online]. Available from: http://www.dtic.mil/docs/citations/ADA425495 (Accessed 17 August 2017).

Lagos, N. et al. (2015) 'On the preservation of evolving digital content–the continuum approach and relevant metadata models', in *Research Conference on Metadata and Semantics Research*. 2015 Springer. pp. 15–26.

Lagoze, C. et al. (2005) Fedora: an architecture for complex objects and their relationships. *International Journal on Digital Libraries*. 6 (2), 124–138.

Lagoze, C. & Van de Sompel, H. (2015) *The Open Archives Initiative Protocol for Metadata Harvesting* [online]. Available from: http://www.openarchives.org/OAI/openarchivesprotocol.html.

Library of Congress (2008) *MARC 21 LITE Bibliographic Format* [online]. Available from: http://www.loc.gov/marc/bibliographic/lite/ (Accessed 26 April 2017).

Library of Congress (2012) *Dublin Core Metadata Element Set Mapping to MODS Version 3* [online]. Available from: http://www.loc.gov/standards/mods/dcsimple-mods.html (Accessed 24 April 2017).

Library of Congress (2015) *PREMIS Data Dictionary for Preservation Metadata, Version 3.0* [online]. Available from: http://www.loc.gov/standards/premis/v3/ (Accessed 17 August 2017).

Library of Congress (2017) *Encoded Archival Description Tag Library - Version EAD3* [online]. Available from: https://www.loc.gov/ead/EAD3taglib/index.html (Accessed 5 April 2017).

Littman, J. (2006) A technical approach and distributed model for validation of digital objects. *D-Lib Magazine*. 12 (5) [online]. Available from: http://www.dlib.org/dlib/may06/littman/05littman.html (Accessed 17 August 2017).

Lubell, J. (2001) Architectures in an XML world. *Markup Languages Theory and Practice*. 3 (4), 399–410.

Maslov, A. et al. (2010) Adding OAI-ORE support to repository platforms. *Journal of Digital Information*. 11 (1) [online]. Available from: http://journals.tdl.org/jodi/article/view/749/640 (Accessed 31 May 2011).

McDonough, J. (2006) METS: standardized encoding for digital library objects. *International Journal on Digital Libraries*. 6 (2), 148–158.

McDonough, J. (2008) 'Structural metadata and the social limitation of interoperability: a sociotechnical view of XML and digital library standards development', in *Balisage: The Markup Conference 2008: Proceedings*. 2008 [online]. Available from: http://www.balisage.net/Proceedings/vol1/html/McDonough01/BalisageVol1-McDonough01.html (Accessed 10 July 2012).

McDonough, J. (2009a) 'Aligning METS with the OAI-ORE data model', in *JCDL '09 Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*. 2009 New York: Association for Computing Machinery. pp. 323–329.

McDonough, J. (2009b) XML, interoperability and the social construction of markup languages: the library example. *Digital Humanities Quarterly*. 3 (3) [online]. Available from: http://digitalhumanities.org/dhq/vol/3/3/000064/000064.html (Accessed 19 April 2017).

METS Editorial Board (2010) *Reimagining METS: an exploration: draft, draft, draft for discussion at Fall 2010 DLF* [online]. Available from: https://www.diglib.org/wp-content/uploads/2011/01/METSNextGeneration.pdf (Accessed 25 March 2011).

METS Editorial Board (2011) *METS 2.0 Data Model Presentation* [online]. Available from: https://www.loc.gov/standards/mets/presentations/METS2-Tom.pptx (Accessed 7 April 2017).

MICE Project (2011) *Measuring Impact Under CERIF (MICE)* [online]. Available from: http://mice.cerch.kcl.ac.uk/ (Accessed 20 October 2011).

Myrick, L. (2004) *Case study: Using METS as a DIP to navigate archived websites* [online]. Available from: http://www.loc.gov/standards/mets/presentations/od3/myrick.ppt (Accessed 5 October 2017).

National Demonstration Test Catchment Network (2011) *National Demonstration Test Catchment Network* [online]. Available from: http://www.demonstratingcatchmentmanagement.net/ (Accessed 22 November 2011).

Neuroth, H. & Koch, T. (2001) 'Metadata mapping and application profiles. Approaches to providing the cross-searching of heterogeneous resources in the EU project Renardus', in *International Conference on Dublin Core and Metadata Applications*. 2001 pp. 122–129.

Nguyen, H.-Q. et al. (2009) 'Schema mediation for heterogeneous XML schema sources', in *Advanced Information Networking and Applications Workshops, 2009. WAINA'09. International Conference on*. 2009 IEEE. pp. 316–321.

Nguyen, H.-Q. et al. (2011) Double-layered schema integration of heterogeneous XML sources. *Journal of Systems and Software*. 84 (1), 63–76.

Open Archives Initiative (2008) *ORE User Guide - Primer* [online]. Available from: https://www.openarchives.org/ore/1.0/primer/ (Accessed 31 March 2017).

Open Archives Initiative (2011) *Open Archives Initiative Protocol - Object Exchange and Reuse* [online]. Available from: http://www.openarchives.org/ore/ (Accessed 12 December 2011).

Parekh, J. et al. (2006) Retrofitting autonomic capabilities onto legacy systems. *Cluster Computing*. 9 (2), 141–159.

Patel, M. et al. (2005) *Semantic interoperability in digital library systems* [online]. Available from: http://opus.bath.ac.uk/23606/1/SI_in_DLs.pdf (Accessed 24 April 2017).

PERICLES project (2017) *PERICLES project* [online]. Available from: http://www.pericles-project.eu/ (Accessed 26 June 2017).

Pollock, R. (2011) *Building the (open) data ecosystem* [online]. Available from: https://blog.okfn.org/2011/03/31/building-the-open-data-ecosystem/ (Accessed 17 August 2017).

Prom, C. J. (2002) Does EAD play well with other metadata standards? Searching and retrieving EAD using the OAI protocols. *Journal of Archival Organization*. 1 (3), 51–72.

Prom, C. J. & Habing, T. G. (2002) 'Using the open archives initiative protocols with EAD', in *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*. 2002 ACM. pp. 171–180.

Pournaras, E. & Miah, S. J. (2012) 'From metaphor towards paradigm—A computing roadmap of digital ecosystems', in *Digital Ecosystems Technologies (DEST), 2012 6th IEEE International Conference on*. 2012 IEEE. pp. 1–6.

Russell, R. (2011) *Research Information Management in the UK: current initiatives using CERIF* [online]. Available from: http://www.ukoln.ac.uk/rim/dissemination/2011/rim-cerif-uk.pdf (Accessed 26 May 2017).

Saleem, K. et al. (2008) PORSCHE: Performance oriented schema mediation. *Information Systems*. 33 (7), 637–657.

Schmidt, D. (2014) Towards an interoperable digital scholarly edition. *Journal of the Text Encoding Initiative*. (7) [online]. Available from: https://jtei.revues.org/979 (Accessed 19 April 2017).

Semantic World (n.d.) *Semantic World - Glossary* [online]. Available from: http://yehuditcohen.com/semWorld/glossary.htm (Accessed 24 April 2017).

Shaker, R. et al. (2002) 'A rule driven bi-directional translation system for remapping queries and result sets between a mediated schema and heterogeneous data sources.', in *Proceedings of the AMIA Symposium*. 2002 American Medical Informatics Association. pp. 692–696. [online]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2244455/pdf/procamiasymp00001-0733.pdf (Accessed 27 April 2017).

Sharma, R. (2007) *Fedora interoperability review* [online]. Available from: http://wwwcache1.kcl.ac.uk/content/1/c6/04/55/46/fedora-report-v1.pdf (Accessed 15 June 2011).

Shaw, E. J. (2001) Rethinking EAD: balancing flexibility and interoperability. *New Review of Information Networking*. 7 (1), 117–131.

Simons, G. F. (1998) 'Using architectural processing to derive small, problem-specific XML applications from large, widely-used SGML applications', in *Summer Institute of Linguistics Electronic Working Papers*. 1998 Summer Institute of Linguistics [online]. Available from: http://www.silinternational.org/silewp/1998/006/SILEWP1998-006.html (Accessed 1 February 2011).

Simons, G. F. (1999) Using architectural forms to map TEI data into an object-oriented database. *Computers and the Humanities*. 33 (1–2), 85–101.

Society of American Archivists (2014) *EAD - Technical Considerations* [online]. Available from: http://www2.archivists.org/sites/all/files/EADRevisionTechnicalConsiderations_0.pdf (Accessed 22 April 2014).

Stein, R. & Coburn, E. (2008) 'CDWA Lite and museumdat: new developments in metadata standards for cultural heritage information', in *Proceedings of the 2008 Annual Conference of CIDOC*. 2008 pp. 15–18.

Taylor, A. G. (2004) *The organization of information*. 2nd edition. Westport: Libraries Unlimited.

The National Archives (2015) *Discovery | The National Archives* [online]. Available from: http://discovery.nationalarchives.gov.uk/ (Accessed 28 August 2015).

UK Government (2006) *e-Government metadata standard v3.1* [online]. Available from: http://www.nationalarchives.gov.uk/documents/information-management/egms-metadata-standard.pdf (Accessed 21 April 2017).

UKOLN (2000) *DESIRE Registry Data Model* [online]. Available from: http://www.ukoln.ac.uk/metadata/desire/registry/docs/datamodel.html (Accessed 21 April 2017).

Unsworth, J. & Sandore, B. (2010) *ECHO DEPository-Phase 2: 2008-2010 Final Report of Project Activities*. [online]. Available from: http://hdl.handle.net/2142/42572 (Accessed 17 August 2017).

Van der Vlist, E. (2007) *Schematron*. Sebastapol, CA: O'Reilly Media, Inc.

Van Zundert, J. (2012) If you build it, will we come? Large scale digital infrastructures as a dead end for digital humanities. *Historical Social Research/Historische Sozialforschung*. 37 (3), 165–186.

Veltman, K. (2001) Syntactic and semantic interoperability: new approaches to knowledge and the semantic web. *New Review of Information Networking*. 7 (1), 159–183.

Waddington, S. et al. (2016) 'PERICLES–Digital Preservation through Management of Change in Evolving Ecosystems', in *European Project Space (EPS) event organized in Colmar, July 2015, associated with the set of conferences ICETE (12th International Joint Conference on e-Business and Telecommunications), ICSOFT (10th International Joint Conference on Software Technologies), SIMULTECH (5th International Conference on Simulation and Modeling Methodologies, Technologies and Applications) and DATA (4th International Conference on Data Management Technologies and Applications).* 2016 pp. 51–74.