**Intermediary XML schemas: constraint, templating and interoperability in complex environments**

**Author:** Richard Gartner

**Address:** Warburg Institute, Woburn Square, London, WC1B 0AH, UK

**Email**: richard.gartner@sas.ac.uk

**Biographical note:** Richard Gartner is a practising librarian and an academic specializing in the theory and practice of digital librarianship. He is currently the Digital Librarian at the Warburg Institute in London. His areas of research have concentrated on the creation and curation of digital collections, digital library metadata and the role of digital libraries in scholarship. He also has an active research interest in the potential of digital libraries for enhancing international development.

**Abstract**

This article introduces the methodology of intermediary schemas for complex metadata environments. Metadata in instances conforming to these is not generally intended for dissemination but must usually be transformed by XSLT transformations to generate instances conforming to the referent schemas to which they mediate. The methodology is designed to enhance the interoperability of complex metadata within XML architectures.

This methodology incorporates three subsidiary methods: these are project-specific schemas which represent constrained mediators to over-complex or over-flexible referents *(Method 1)*, templates or conceptual maps from which instances may be generated *(Method 2)* and serialized maps of instances conforming to their referent schemas *(Method 3)*.

The three methods are detailed and their applications to current research in digital ecosystems, archival description and digital asset management and preservation are examined. A possible synthesis of the three is also proposed in order to enable the methodology to operate within a single schema, the Metadata Encoding and Transmission Standard (METS).

**Keywords:** XML, intermediary XML schemas, metadata, interoperability, digital asset management, digital preservation, METS (Metadata Encoding and Transmission Standard)

**Introduction**

In general metadata schemes are designed to act as carriers for metadata in the form in which it is to be disseminated. In a number of cases, however, it may be useful to devise schemes which act as mediators to others, in which their constituent metadata must undergo some degree of transformation before it is delivered to users. There may be several reasons why this circuitous approach may be valid, particularly if the final scheme into which metadata is converted presents problems of over-complexity or limited interoperability. We may call these schemes ‘intermediary’ in the sense that their metadata acts as a mediator to its equivalent in a more established scheme, which we may refer to as its ‘referent’.

This article examines XML (eXtensible Markup Language) metadata schemas which function as intermediaries in this way. Such schemas perform this function either to other ‘referent’ schemas or sometimes to alternatively-realized instances conforming to themselves. When used in this way, their constituent metadata must be converted by XSLT (eXtensible Stylesheet Language – Transformations) transformations into instances which can be validated against their referents. It is in this finalized form that this metadata is then disseminated to the end user.

These ‘intermediary XML schemas’ (Gartner, 2018) offer the potential to alleviate some significant problems which have been noted in employing XML-based metadata structures in digital asset management: these include the over-rigidity of their hierarchical structures, their lack of flexibility and the impediments that they may present to discovery. Conversely some established XML schemas may also cause problems because they are too flexible in their design, offering multiple approaches to encoding which can create problems for system design.

RDF-based (Resource Description Framework) approaches to the representation of metadata, such as those employed by the Fedora Content Model Architecture (CMA), are occasionally cited as alternatives which can obviate these problems (for instance by Lagoze et al. (2005) and Johnson and Dehmlow (2019)) and for enhanced metadata openness, serendipity and facet-based navigation (Alemu et al., 2012). They have, however, been criticized, particularly within the digital asset management and library communities, as viable strategies for metadata management. A highly negative report from the UK library community, for instance, concluded that ‘we have yet to see any real examples of benefit emerging from …[Linked Open Data (LOD)] projects in this area, or elsewhere’ (Hawtin et al., 2011, p.26), citing issues of the complexity of data modelling and ontology development (p.17) and problems of data cleansing (p.17). Because one of the key rationales of RDF-based LOD is its ability to interoperate across silos of metadata, particular problems arise for digital preservation as its fluid boundaries are disjoint with the package-based models of OAIS (Open Archival Information Systems) on which much practice is built (Gartner, 2018, p.71) and also make it more difficult to establish spheres of responsibility for preservation (Gartner, 2016, p.92).

This article introduces three methods which may be termed ‘intermediary XML schemas’ and which attempt in their various ways to reconcile flexibility with the conflicting requirements of interoperability to which such flexibility can act as an impediment. In doing so it aims to enhance the potential of this architecture within digital asset management as a viable alternative to RDF.

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

**Background: literature review**

Interoperability in metadata has been defined as the ability to exchange data such that it can be used in diverse software applications without modification (Schmidt, 2014) or without manipulation in a new system (Taylor, 2004, p. 169). Ideally no human intervention should be required in these processes (Baumann, 2011). To do this requires a degree of negotiation to ensure shared semantics (Veltman, 2001, p.167; Gartner, 2016, p.23). As such, it is more difficult to achieve than interchange which does require manipulation of data to allow its incorporation into a receiving system.

Despite its credentials as an interchange format, XML has been criticized for limits to its potential for interoperability. Gradmann, for instance, points out that it is interoperable at the syntactic and structural levels but not at the semantic (Gradmann, 2010, p.159) and so standard validation techniques test conformance to syntax and structure alone (Jacinto et al., 2004, p.2; Decker et al., 2000, p.67; Gartner, 2018, p.24). Even syntactic and structural interoperability may be limited by over-flexible, extensible, modular and abstract schemas (McDonough, 2008). The validation of sematic interoperability can usually only be achieved by invoking additional validation techniques such as XCSL (XML Constraint Specification Language) (Jacinto et al., 2004) and Schematron (Van der Vlist, 2007; Gartner, 2018, p.25).

To overcome some of the limitations of the XML syntax, a number of projects have employed what they have termed ‘mediating’, ‘intermediate’ or ‘intermediary’ schemas in order to mediate to others in various ways. In a few cases, synthetic schemas have been devised which act as receptacles into which other schemas can be merged in order enable the integration of their constituent metadata. Almarimi and Pokorny, for instance, employed a common ‘global’ schema of this kind, created by the merger of several source schemas, as a core part of a complex data integration system (2005); similarly merged schemas were created by the PORSCHE (Performance Orientated Schema Mediation) (Saleem et al., 2008), XMediator and XINTOR (Nguyen et al., 2009; Nguyen et al., 2011) projects (Gartner, 2018, p.45).

Other projects defined common intermediate schemas to facilitate data integration from heterogeneous sources: the National Institute of Health, for instance, produced such a schema to merge XML query results (Shaker et al., 2002) and the Xyleme project devised one to facilitate views across diverse XML data sources (Abiteboul et al., 2002). A further application utilized a mediating schema generated from XML source instances to facilitate the definition of an OWL (Web Ontology Language) ontology (Bohring et al., 2005) (Gartner, 2018, p.46).

These intermediary schemas essentially function as mapping mechanisms: they are more than simple ‘crosswalks’, one-to-one ‘semantic and/or technical mapping[s] (sometimes both) of one metadata framework to another metadata framework’ (Patel et al., 2005, p.21), such as Dublin Core to MODS (Metadata Object Description Schema) (Library of Congress, 2012) or EAD (Encoded Archival Description) to MODS (Bountouri and Gergatsoulis, 2009). In terms of Boley and Chang’s often-cited definitions of techniques for schema-level interoperability (Boley and Chang, 2007) these examples are ‘switching’ schemas, which are devised to synthesise diverse data sources: a well-known schema of this kind is OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) which employs Dublin Core elements to provide this function (Lagoze and Van de Sompel, 2015).

**Proposed methods**

As the above literature review indicates, the terms ‘intermediary’, ‘mediating’ or ‘intermediate’ schemas have been employed in a relatively narrow sense as means to allow the synthesis of diverse data from heterogeneous sources. The methods described in this article attempt to extend their usage beyond this function into such areas as the reconciliation of divergent architectures, the templating of complex data objects and the definition of abstract maps of complex metadata environments which are designed to enable greater interoperability between their constituents.

***Method 1*: constrained, project-specific intermediary schemas**

The first intermediary schema method, referred to as *Method 1* throughout this article, is that of project-specific XML schemas which represent an intermediary to an established counterpart in common use in a given user community. This latter schema, the ‘referent’ of the intermediary, may be widely established but nevertheless problematic in its application within the context of a given project and its requirements.

Two well-established standards in their respective user communities present these problems. In the area of research information management CERIF (Common European Research Information Format) (European Organisation for International Research Information, 2012) has established itself as the *lingua franca* for the interchange of this data. At least initially, however, it proved itself problematic as it consisted of 192 separate schemas linked together by a complex system of identifiers which emulated its origins as relational SQL tables (EuroCRIS 2012a). The complexity of encoding that resulted from this was exacerbated by its reliance on external semantic schemes to provide the linkages between these schemas. Later versions of CERIF (from version 1.4) (EuroCRIS 2012b) rationalized this structure into a single schema.

The archival standard EAD (Library of Congress, 2017) has also been criticized in the literature for its limited interoperable potential owing to its document-centric design (Dow, 2009, p.111) and the multiplicity of options that it allows for the expression of a single concept (Shaw, 2001, p.123); these can lead to inconsistencies in markup which can limit its potential as a discovery medium (Prom, 2002) and make the design of systems based on this standard more difficult (Blanke and Kristel, 2013).

*Method* 1 was first proposed as a means to ameliorate these problems in an article in 2011 which introduced CERIF4REF, an ‘intermediary schema’ to CERIF (Gartner, 2011). This schema was designed to offer a heavily-constrained view of a CERIF application in which the network of linkages necessary to encode concepts in the referent schemas were hard coded into the semantic definitions of their intermediary. This greatly simplified the overall architecture of an application by reducing the complex web of separate schemas into a single unified structure. But it also offered opportunities to reconcile divergent architectures between CERIF and other schemas to which its instances needed conversion.

One such schema was devised for the 2008 Research Assessment Exercise (RAE) in the UK, a periodic exercise in which higher education institutions submit research outputs for assessment by expert panels, the outcome of which determines central government funding for their ongoing research activities. As part of this assessment, information on research activities could be submitted as XML conforming to a bespoke schema called RAE. This had several notable architectural incongruities with CERIF, notably in its approach to the aggregation of data: to record the research assistants attached to a project, for instance, CERIF would list all explicitly while RAE would record a count of their numbers (Gartner, 2018, p.28).

To alleviate these problems, CERIF4REF attempts to extend the technique of architectural processing which had earlier been developed in SGML (Standard Generalized Markup Language), the predecessor to XML: this allows the construction of constrained DTDs (Document Type Definition) for specific applications, the elements and attributes of which are mapped to their equivalents in others which are problematic because of their over-flexibility (such as the TEI (Text Encoding Initiative) (Simons, 1998)). In architectural processing, the serialization of an architecturally mapped DTD to its established equivalent is carried out by parsers during the process of validation but in the case of an intermediary schema such as CERIF4REF this is achieved through an XSLT transformation.

Utilising XSLT allows the realization of intermediary schemas to be accomplished with greater flexibility than the mapping of elements and attributes alone. Data may be aggregated during the XSLT process itself, so allowing CERIF4REF to contain disaggregated data in a manner akin to CERIF but to generate its aggregated form as required by RAE. Stronger validation of semantic links may be achieved by using ID/IDREF pairs to link components in CERIF4REF instead of the use of identifiers from external semantic schemes: these links are checked by standard XML validation parsers without having to invoke external programs (such as Schematron) which would be needed to validate the linkages between schemas in CERIF itself. Any semantics that are required by CERIF but are not part of the metadata recorded in instances (such as the identity of the semantic schemes that define the linkages between schemas) may be hard-coded into the XSLT stylesheets instead of the CERIF4REF schema itself, so much simplifying its architecture and that of its instances (Gartner, 2018, p.28).

The overall architecture of CERIF4REF and its relationships to CERIF and RAE is shown in Figure 1.

<FIGURE 1 HERE>

Figure 1 CERIF4REF and its relationships to RAE and CERIF

A further *Method 1-*type intermediary schema, which addresses other issues than over-complexity and incongruent architectures, is the CENDARI (Collaborative European Digital Archive Infrastructure) Collection Schema (CCS) (Gartner, 2015). This is designed to act as an intermediary to EAD in order to alleviate some of the problems already noted with this widely-used community standard. CCS is a more data-centric schema than its referent, omitting many of the latter’s more text-centred elements which derive from its origins in the traditional archival finding aid. It therefore generates a limited EAD instance only but allows more readily interchangeability between systems while still ensuring compatibility with the community-established standard (Gartner, 2018, p.30).

This method has significant potential for simplifying the implementation of any complex XML schema and enhancing its interoperability. The recently defined Europeana Data Model (EDM) (Europeana, 2016), for instance, is available as 27 interlinked XML schemas which owe their origins to the numerous RDF-based semantic schemes which have been incorporated into it. As is the case with CERIF, the management and validation of these is cumbersome and requires the use of *Schematron* to be achieved effectively. Employing an intermediary schema akin to CERIF4REF could facilitate the application of this important data model considerably.

There are, however, a number of limitations to this method which mitigate against its undoubted potential. An intermediary schema may easily become unsynchronized from the XSLT stylesheets required for its realization, in a way that was not possible in architectural processing where the mapping between a DTD and its referent is integrated into its definition. More seriously, the adoption of this approach may produce a proliferation of bespoke schemas which are not conformant to published standards, and so produce a messy metadata environment which could impede overall interoperability. If these schemas, rather than their referents, form the basis on which systems are designed, the possibility of a landscape of isolated applications, unable to communicate with each other, becomes more rather than less likely (Gartner, 2018, p.57).

***Method 2*: intermediary schemas for templates or conceptual modelling**

A further use of intermediary schemas, referred to as *Method 2* in this article, is as templating mechanisms for the delivery of complex digital objects. As is the case with the schemas described as *Method 1* above, these rely on XSLT transformations for the realization of instances and their consequent dissemination, but do not act as intermediaries to other schemas: instead, they encode skeletal templates from which instances conformant to themselves are generated (Gartner, 2018, p.34).

Two articles in the literature (Gartner, 2012 and Gartner, 2014) discuss the use of the METS (Metadata Encoding and Transmission Standard) standard as an intermediary schema of this type. In the first of these, METS is used as a templating mechanism for the delivery of complex biological nanoimaging images. These must be constructed by the combination of their component files in ways which differ according to the experiment in which they originate (Gartner, 2012a, p.25): one, for instance, requires the combination of three raw images and a further ‘background subtraction’ image which are then processed by a specific software package to produce a TIFF image file for delivery.

The intermediary schema method deployed here separates the content of each realized image from its underlying structure. This structure is encoded within the METS structural map, a separate one for each experiment and its required combination of components. Instead of following the normal procedure within METS of referencing the component files of the compound object within the METS file section, empty placeholders mark their place there. When the intermediary METS file is realized by the application of its associated XSLT transformation, these placeholder elements are populated with references to the component files required for each object: these components are themselves listed in small subsidiary METS files (Gartner, 2012a, p.31).

Separating the structure of a complex compound object from its constituent components in this way allows the latter to be reused in multiple objects, each time recombined as required. In this way, this method emulates one of the core advantages often claimed for RDF-based models, such as the Fedora CMA, for the encoding of complex objects. This flexibility of modelling at multiple levels of granularity is achieved entirely within an XML architecture, overcoming some of the criticisms that have been made of over-rigidity and impediments to discovery that some claim are inherent to this syntax (Han, 2006, p.236).

Figure 2 illustrates the overall architecture of this method.

<FIGURE 2 HERE>

Figure 2 Method 2 overall architecture

The second article that discusses this method (Gartner, 2014) extends the templating function of METS to multiple levels of granularity within an integrated architecture for a digital archive. In this case, a collection of complex and heterogeneous environmental data, compiled from multiple sources, is structured by a series of METS intermediary schemas each of which represents a separate level of granularity within the whole architecture. Each data type, for instance flow gauge or weather station data, is defined structurally within a METS file as in Gartner (2012). Each object’s metadata is stored within a small, subsidiary METS file and is used to populate the higher-level template file on delivery. Above the level of these templates a further level is used to structure the archive as a whole, once again using the METS structural map to do so.

The structural map in this top-level METS file is populated by <mptr> elements: the <mptr> (METS Pointer) element signifies that the content of an external METS file should fill a <div> (division) within this file. . It is here that the secondary-level templating METS files are invoked; these are then populated by the relevant metadata for each object. In this way, the entire archive is defined within a single, integrated architecture but the components with which it is populated can be used and re-used flexibly at any relevant point.

This approach achieves a separation of the structure of a complex digital object from its content in a manner analogous to Fedora Content Models but does so entirely with an XML architecture. The same architecture can encompass multiple levels of granularity from the most abstract to the smallest metadata components. Doing this within XML allows much easier data management than would be possible using RDF-based methods and also ensures compatibility with package-based digital preservation models such as OAIS (Open Archival Information System) (Gartner, 2018, p.61). Utilising the METS standard as the overall framework should, despite some problems of interoperability with this schema (noted by, for instance, Guenther (2008) and McDonough (2006)), allow greater exchangeability and transferability than is possible with Fedora Content Models or similar RDF-based models.

One significant drawback to this method compared to RDF is that the METS structural map records syntactic rather than the semantic relationships which are readily incorporated within the Fedora Content Model. The structural map encodes ‘part-to-whole' or ‘is-constituent-of' relationships within its hierarchies but not core semantic relationships, such has ‘has-equivalent’ or ‘has-annotation’, which are readily handled by Fedora (Gartner, 2018, p.62).

METS does have a feature which can alleviate this semantic deficit to some extent. Its structural linking facility, the <structLink> element, allows linkages which cut across the hierarchies of the structural map. These may be used, for instance, to express hyperlinks within a website archived in the METS architecture (Myrick, 2004). This element may have an attribute ARCROLE which indicates the role of a link and can contain a URI (Uniform Resource Identifier) expressing a semantic relationship, akin to the use of URIs to express the predicate of an RDF triple. In this way, any semantic relationship that can be expressed in Fedora or other RDF-based architectures can be incorporated into METS-based intermediary schemas of this kind (Gartner, 2018, p.77).

***Method 3*: abstracting and mapping XML instances**

A third type of intermediary schema, referred to as *Method 3* in this article, attempts to enforce the constraining functions of *Method 1* without the need for generating bespoke, project-specific schemas. Instead of designing a constrained schema from which instances conforming to its referent can be generated by XSLT, this method uses the METS structural map to serialize the structures underlying the referent’s instances. This structural mapping can then be used to enhance the interoperability of the referent schema by facilitating the exchange of metadata conforming to it and to generate alternative views of this metadata by XSLT transformations (Gartner, 2012b).

The METS structural map here employs a ‘striped’ syntax, originally proposed by (Habing and Cole, 2009) as a way of serializing OAI-ORE (Open Archives Initiative Object Reuse and Exchange) aggregations in the METS structural map. Three nested <div> (division) elements emulate the structure of an RDF triple: the outermost of these <div>s represents the subject of this ‘triple’, the second its predicate and the third its object. Each <div> references a component in the instance metadata: this is achieved by employing XPointer references, a standard mechanism within XML for establishing linkages within and between XML instances.

Figure 3 illustrates how an association expressed in CERIF between a person and an article which they have authored may be represented in this way. In CERIF this relationship requires the linking of three tables, one identifying the person (labelled ‘Person details’ in the diagram), one the article (‘Research publication details’) and one the relationship (‘Class details’) between the two. This is expressed in the structural map by the tripartite structure of three nested <div> elements (marked a in the diagram). Each of these elements contains an <area> attribute which is used within METS to delineate a section of an external file referenced by the <div>: the BEGIN attribute contains the XPOINTER reference to the component in the referent instance with which the <div> should be populated.

<FIGURE 3 HERE>

Figure 3 Using striped syntax to serialize linkages in CERIF

These tripartite <div> structures form the lowest-level granular components of the structure that is encoded in this way. Above it the full range of internal relationships that can be expressed within the METS architecture can be accommodated: this can include the hierarchies of the structural map itself and the cross-cutting linkages that can be expressed using the <structLink> structural link elements. A rich set of relationships can therefore be expressed which can accommodate the requirements of any application.

This method attempts to perform a similar function to the bespoke intermediary schemas of *Method 1*: it attempts to enhance interoperability by introducing a constrained representation of the structures of their instances which is tailored to the requirements of a given application. As is the case with the schemas of *Method* 1, this representation encodes a selective view of the range of possible content that instances conforming to their referent schema can include; it can then act as a template for navigating the semantic structures of these instances so rendering them more interoperable and an easier basis on which to build application than the referent schema (Gartner, 2018, p.33).

Unlike *Method 1*, however, it does this within a single architecture, that of the METS structural map. It therefore avoids the potential problems of an increasingly messy overall metadata environment in which a large number of bespoke intermediary schema may proliferate (Gartner, 2018, p.58). It should also provide enhanced archival robustness compared to such schemas as CERIF4REF or CCS as the METS standard is itself designed to function as a packaging mechanism within the OAIS reference model (Gartner, 2018, p.59).

The overall architecture of Method 3 is shown in Figure 4.

<FIGURE 4 HERE>

Figure 4 Overall architecture of Method 3

Using METS instead of designing project-specific schemas is not without its problems. The architecture of its structural map is relatively rigid in its nested hierarchies, although this can be mitigated, as noted above, by the employment of cross-cutting structural link elements. This rigidity can be an enhancement to its potential for enhancing interoperability but may be more limiting than the possibilities offered by the design of a bespoke intermediary schema (Gartner, 2018, p.60).

A further problem with this method may be maintaining synchronization between the schema and the structures of the referent schema instances: any change to the latter (for instance if the instances employ some of the alternative encoding possibilities offered by many schemas) would break the XPointer links and render the abstracted mapping in the METS file inoperable. Considerable care will be required to detect such breakages so that synchronization can be maintained.

**Potential applications**

The three methods described above have multiple applications and have already been employed with some success. The bespoke schemas of *Method* 1 were used in a research information management project, Readiness for REF, which examined ways in which higher education institutions in the UK could submit data to that country’s 2014 REF (Research Excellence Framework) exercise (Centre for e-Research, 2011). It was used specifically as the basis of plugins to institutional repository systems such as e-Prints in order to enable data from these to be accessed and shared (Cox, 2011). It also readily incorporated later modifications to the CERIF standard to incorporate impact indicators (Gartner et al., 2013).

*Method 2* proved itself viable as an alternative to the Fedora Content Model Architecture for encoding templates and higher-level conceptual models in the projects described in Gartner (2012) and Gartner (2014) although it was not ultimately implemented in these projects because the Fedora model was the established infrastructure within which they were required to operate when they were planned. *Method 3* proved itself capable of attaching contextual information to research outputs from which metadata could be output in any required format including CERIF4REF (Gartner, 2018, p.58).

In addition to these past projects, there are three areas of contemporary research in which such schemas may prove relevant and useful. These are 1) digital ecosystems, particularly as applied to research environments in the digital humanities, 2) enchancing archival description to facilitate the integration of diverse metadata sources and 3) enhancing the viability of XML in digital asset management and preservation as an alternative to RDF.

 The first of these is that of the digital ecosystem, a model of information environment in which open communities of users resolve problems collectively without centralized control (Boley and Chang, 2007, pp.1–4). Such ecosystems are variously characterized in the literature as being cyclical in nature (Pollock, 2011), self-regulating, continuously developing (Hitruhhina, 2012), self-organizing and scalable (Briscoe and De Wilde, 2006, p.17; Gartner, 2018, p.65).

The digital ecosystem metaphor is commonly applied to research environments (Pournaras and Miah, 2012), particularly in the digital humanities where ‘change, collaboration and engagement’ are considered crucial (Anderson, 2013, p.20) and communities of researchers rather than technologies are the key determinants of the forms that they take (Anderson and Blanke, 2012, p.161). One component that is considered anathema to certain advocates of the digital ecosystem is the notion of their being constrained by standards, including those of metadata: Van Zundert, for instance, denigrates these as like to impede innovation and research (Van Zundert, 2012, pp.173–174). To abandon metadata standards in particular is highly problematic as it risks reduced interoperability, isolation, the creation of discrete silos of research infrastructures and often the reinvention of techniques and methodologies which have already proved viable (Gartner, 2018, p.66).

One way in which intermediary schemas, particularly of *Method 1*, could alleviate this is to form the cores on which research infrastructures and digital ecosystems are built. An article from 2013 on the CENDARI project proposed using intermediary schemas as the base infrastructure of a research environment and supplementing these with ontologies which change to reflect ongoing research that derives from the archival materials described in these schemas (Gartner and Hedges, 2013, pp.63–64). Using an intermediary schema ensures that the infrastructure would be grounded in established community practice (such as the EAD standard for archives) and so would not be isolated from previous research, whilst the changing ontologies would allow flexibility and avoid any impediment to new developments and innovation (Gartner, 2018, p.68).

A further area in which intermediary schemas of this type offer considerable potential is that of enhancing archival description to allow information on resources from diverse sources to be integrated more readily than is currently allowed by EAD. Some current projects, most notably the European Holocaust Research Infrastructure (EHRI) (EHRI Project, 2017), attempt to integrate the records of hundreds of repositories to create something approaching a union catalogue but are frequently hindered by the inconsistencies and flexibilities of EAD noted earlier. To overcome these problems, EHRI has adopted the use of graph-based NoSQL databases which render the cross-searching of archival records easier (Blanke and Kristel, 2013) but at the cost of some uncertainty about the possibility of ensuring the persistence of research data in the long term (Blanke et al., 2017, pp.19–20) as this data is embedded in complex, bespoke database structures far removed from standard archival practice (Gartner, 2018, p.69).

*Method 1* intermediary schemas allow the incorporation of more data-like components as mediators to document-centric schemas such as EAD: it therefore allows them to function in a manner akin to the graph-like databases of NoSQL while retaining the advantages of the XML syntax. They do so without breaking the link to EAD and the large community of archival practice on which it is based. They could therefore be particularly valuable for a project such as EHRI in which the requirements of functionality and long-term persistence are rendered at odds by the adoption of database methodologies.

The third area of research to which this methodology has considerable potential is that of digital asset management and preservation metadata in general and as an alternative to linked open data (LOD) in particular. Although it forms the basis of such widely-respected repository systems as Fedora Commons, linked open data has, as discussed earlier, been much criticized as a viable format for metadata in digital asset management and preservation.

Intermediary schemas, particularly *Methods 2* and *3*, offer potential routes to resolving some of the problematic features of RDF as they can emulate many of the features of its approaches to metadata wholly within an XML environment. In particular, they allow the modelling of complex metadata structures within the discretely packaged METS structure so obviating many of the problems of fluid boundaries presented by linked open data. They also ensure greater compatibility with OAIS as METS is designed to operate as a Submission Information Package (SIP), Archival Information Package (AIP) and Dissemination Information Package (DIP) within the OAIS conceptual model (Gartner, 2018, p.72). This may also make them a viable alternative to JSON-LD (JavaScript Object Notation for Linked Data) methods for encoding such metadata as IIIF (International Image Interoperability Framework) (IIIF Consorium, 2019) manifests.

This methodology is also relevant to current research into digital preservation which seeks to abandon the custodial model of OAIS in favour of more process-based models where there is not necessarily a definitive final state for an archived object. The recent PERICLES (Promoting and Enhancing Reuse of Information throughout the Content Lifecycle taking account of Evolving Semantics) project, for instance, rejects distinctions between the active and archival phases of a digital object’s lifecycle (Waddington et al., 2016, p.54), and instead employs the model of a digital ecosystem to support its model of continuous change in complex objects such as software-based art (Lagos et al., 2015, p.18; Gartner, 2018, p.72).

The hybrid digital ecosystem proposed by Gartner and Hedges (2013) may prove a useful model for reconciling the requirements of dynamic digital objects and standards-based digital preservation practice in projects such as PERICLES. OAIS-compatible intermediary schemas encoded in METS could, for instance, record the stable core of significant properties for these objects and be supplemented by ontologies to encode the dynamic properties which change as their lifecycles develop. It may also be possible to use the serialized maps of *Method 3* to fulfil the functions of ontologies, so ensuring a fully XML environment; this, however, will require considerable further research to develop to a viable stage (Gartner, 2018, p.72).

**A possible synthesis of the three methods**

Although the three methods described here, which might all be legitimately described as ‘intermediary XML schemas’, have already proved useful and have the potential for enhancing current research in a number of areas, there exists the possibility of synthesizing the three into an overall methodology which would enhance their potential for enabling greater interoperability within XML architectures. In particular, it should be possible to bring together the templating and conceptual modelling of *Method 2* and the serialized maps of *Method 3* in a way which could fulfil many of the functions of *Method 1* without the risk, highlighted earlier, of creating messy overall environments composed of numerous bespoke, project-specific schemas.

One way in which this could be done is employing the serialized maps of *Method 3* to facilitate the ingest of metadata from diverse sources, each map representing an abstraction of the instances from each source. A higher-level map, also presented as a *Method 3*-type METS structural map, could then attempt to consolidate these mappings for a given domain. This would produce a synthesized abstraction of these serialized maps from which template files or conceptual models of the type created by *Method 2* could be generated. These in turn could be used to generate new instances within the domain, instances which would have some degree of interoperability by virtue of their conformance to this multi-level mapping. If necessary, or more convenient, XML schemas could also be generated from this higher-level abstraction of the *Method 3* maps (Gartner, 2018, pp.79–80). Figure 5 illustrates this potential architecture.

<FIGURE 5 HERE>

Figure 5: Model for a potential synthesis of intermediary schema methods (from Gartner, 2018, p.80)

Such a methodology would allow the functionality of the bespoke schemas of *Method 1* to be achieved within the more controlled and standardized environment of the METS structural map. It would do so using a widely-established community-based standard which is compatible with digital preservation standards such as OAIS and would operate wholly within a robust XML architecture.

**Conclusions and future work**

The three intermediary schema methods detailed here have significant potential for enhancing the management and preservation of complex digital objects and their metadata. *Method 1* can do so by constraining complex and over-flexible schemas to enhance interoperability and facilitate the sharing and transmission of their conformant metadata. *Method 2* offers an alternative to RDF-based metadata environments, which present significant problems for digital asset management and preservation, by emulating their templating functionality within robust XML architectures. *Method 3* offers the potential to perform the constraining functions of Method 1 within a standardized environment, the METS structural map and to create a mechanism by which the templates of *Method 2* may be generated (Gartner, 2018, p.82).

The three methods and the proposed synthesis are summarized in the following table:-

<TABLE 1 HERE>

All three may be developed further in future research. Method 1 may be enhanced by examining new ways in which the association between an XML instance and its XSLT transformation may be formalized more robustly (Gartner 2018, p.75). Method 2 may be improved by developing its semantic linkages into forms that are more sophisticated than the relatively basic, primarily syntactic, ones which it presently allows. This may be achieved by making more sophisticated use of METS’s structural linking facilities that the method currently employs (Gartner 2018, p.76). Method 3 requires further work ensure that the semantic maps that it employs remain synchronized to the schemas to which the instances that it references conform. This could possibly be achieved by devising a further level of abstraction to encode these maps; the definition of the syntax for these will require extensive further work (Gartner 2018, p.76).

In their current state, however, all three methods manage to reconcile the divergent requirements of flexibility and interoperability within XML architectures. They also allow the possibility of new and rapidly changing methods (such as ontologies) to be reconciled with existing practices and established standards by acting as mediators to these. Above all, by allowing the functionality of RDF-based approaches to metadata to be emulated within XML, they strengthen the viability of the latter as an infrastructure within the practices of digital asset management and preservation (Gartner, 2018, p.88).

**References**

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

Alemu, G. et al. (2012) Linked Data for libraries: Benefits of a conceptual shift from library-specific record structures to RDF-based data models. *New library world*. 113 (11/12), 549–570.

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*. 3147–164.

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*. 72147–156.

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

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.

Centre for e-Research (2011) *R4R: Readiness for REF* [online]. Available from: http://r4r.cerch.kcl.ac.uk/ (Accessed 21 February 2011).

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).

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

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

European Organisation for International Research Information (2012) *CERIF 1.3* [online]. Available from: http://www.eurocris.org/Index.php?page=CERIF-1.3&t=1 (Accessed 25 January 2012).

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

Gartner, R. (2015) An XML schema for enhancing the semantic interoperability of archival description. *Archival Science*. 15 (3), 295–313.

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.

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 7 July 2012).

Gartner, R. (2018) *Intermediary XML schemas*. PhD Thesis thesis. City, University of London. [online]. Available from: http://openaccess.city.ac.uk/20288/1/Gartner%2C%20Richard.pdf (Accessed 20 September 2018).

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

Gartner, R. (2012a) METS as an ‘intermediary’ schema for a digital library of complex scientific multimedia. *Information Technology and Libraries*. 31 (3), 24–35.

Gartner, R. (2012b) The digital object in context: using CERIF with METS. *Journal of Library Metadata*. 12 (1), 39–51.

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.

Guenther, R. (2008) *Guidelines for using PREMIS with METS for exchange*. [online]. Available from: http://www.loc.gov/standards/premis/guidelines-premismets.pdf (Accessed 30 May 2013).

Habing, T. & Cole, T. (2009) *Candidate approaches for describing ORE Aggregations in METS: Preliminary discussion draft*. [online]. Available from: http://ratri.grainger.uiuc.edu/oremets/ (Accessed 21 February 2014).

Han, Y. (2006) A RDF-based digital library system. *Library hi tech*. 24 (2), 234–240.

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).

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).

IIIF Consorium (2019) *Home — IIIF | International Image Interoperability Framework* [online]. Available from: https://iiif.io/ (Accessed 27 September 2019).

Johnson, D. & Dehmlow, M. (2019) Digital Exhibits to Digital Humanities Expanding the Digital Libraries Portfolio. *New Top Technologies Every Librarian Needs to Know: A LITA Guide*. 123.

Lagos, N. et al. (2015) ‘On the preservation of evolving digital content–the continuum approach and relevant metadata models’, in Emmanouel Garoufallou et al. (eds.) *Research Conference on Metadata and Semantics Research*. [Online]. 2015 Manchester: 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 (2017) *EAD: Encoded Archival Description Official Site* [online]. Available from: http://www.loc.gov/ead/ (Accessed 4 April 2017).

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

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).

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

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.

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).

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.

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.

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

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).

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. p. [online]. Available from: http://www.silinternational.org/silewp/1998/006/SILEWP1998-006.html (Accessed 1 February 2011).

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*. 165–186.

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.

**Figure 1**

****

**Figure 2**



**Figure 3**



**Figure 4**



**Figure 5**



**Table 1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Method 1** | **Method 2** | **Method 3** | **Synthesis** |
| **Schema employed** | Bespoke | METS | METS | METS |
| **Primary function** | Constraint | Templating | Constraint | Constraint |
| **Relation to referent schema** | XSLT transformation | XSLT transformation | Mapping | Mapping |