Submitted on: 28.06.2019

Archives, museums and libraries: breaking the metadata silos

Richard Gartner
Warburg Institute Library, Warburg Institute, London, United Kingdom.
E-mail address: richard.gartner@sas.ac.uk

Raphaële Mouren
Warburg Institute Library, Warburg Institute, London, United Kingdom; Centre Gabriel Naudé, EA 7286.
E-mail address: raphaele.mouren@sas.ac.uk

Abstract:

To overcome impediments to access for the rich collections held in cultural heritage institutions, some alignment of metadata standards and a consequent enhancement of interoperability are essential. The three sectors in which much of our cultural heritage is to be found, libraries, archives and museums, each employ different approaches to metadata which make interoperability between them, and consequently between the constituencies and communities within which they were defined, difficult to achieve.

This paper introduces a metadata strategy devised for the Photographic Collection of the Warburg Institute, London, which aims to break down some of the barriers between these approaches. This strategy employs established XML-based metadata schemas from the digital library community as the basis for establishing interoperable metadata. It uses CIDOC-CRM (the core standard from the museum sector) as a conceptual backbone for metadata structures but then serializes the data model produced into standards from the library sector such as MODS for descriptive metadata and METS for packaging. The resulting metadata can also interoperate to some extent with EAD, the established standard in the archives sector, by utilising existing crosswalks.

This strategy allows the metadata requirements of all three sectors to be accommodated in interoperable, easily-managed schemes. It should allow an important step to be taken towards moving metadata practices within archives, museums and libraries in the same direction.

Keywords: metadata, interoperability, XML CIDOC-CRM, METS, EAD, museums, libraries, archives.

Introduction

Access to the rich collections held in cultural heritage institutions is crucially dependent on the availability of good quality metadata in order to enable their discovery. This *descriptive* metadata is needed to allow potential users to find out the existence of heritage material that might interest them and to assess its potential relevance to their needs. To enable this discovery beyond the boundaries of a single institution, some alignment of metadata standards is essential: this needs to take place at the levels of both semantics (the fields within which the metadata is held) and content rules (the form taken by the metadata that fills these fields).

Substantial strides have been made within the three sectors of libraries, museums and archives to enable some degree of standardization of metadata practice. In libraries, the MARC (Machine Readable Cataloging) standard, first devised in the 1960s, established a uniform practice for the semantics of library catalogue records, and such initiatives as the Anglo-American Cataloging Rules (AACR2) and the Library of Congress Name Authority File (LCNAF) have ensured some degree of standardization of metadata content. This move towards standardization has enabled the easy sharing and transfer of cataloguing records and enabled large union catalogues such as *WorldCat* to operate.

In the archives sector, the Encoded Archival Description (EAD), an XML-based standard for the encoding of archival finding aids first devised in the 1990s, has gone some way towards emulating the approach of MARC. In addition, ISAD-G (General International Standard Archival Description) has provided a standardized scheme for the content of finding aids in a manner analogous to AACR2. Despite a number of well-documented problems which arise from the flexibility of encoding methods allowed by EAD [1], it has had some success in enabling cross-collection union catalogues of archives (such as the UK National Archives *Discovery* initiative (https://discovery.nationalarchives.gov.uk/)) to be constructed.

Within the museums sector, the CIDOC-CRM model, also defined in the 1990s, attempts to define a syntax and formal structure for describing cultural heritage materials in the form of an ontology: this not only defines the classes by which these materials may be described but also the potentially complex web of relationships between these classes in a much more sophisticated manner than is possible in the ‘flat’ files of MARC records. CIDOC-CRM underlies several projects which seek to consolidate holdings from diverse museum collections including the British Museums’s *Researchspace (*https://www.researchspace.org/) and Oxford University’s CLAROS project (http://www.clarosnet.org/XDB/ASP/clarosHome/) .

All three standards have achieved some degree of success within their respective sectors in enabling the sharing and transmission of metadata. There is, however, little interoperability between these three approaches, and consequently between their respective communities, owing to their differing underlying architectures which in turn owe their origins to the very different approaches to metadata that have applied for centuries within each sector. Without this interoperability it is difficult to enable cross-sector or cross-community discoverability, to allow the valuable heritage materials held within, for instance, the library sector to be accessible to users in archives or museums. One of the key challenges in coming years must be to find ways to bridge the gaps between these diverse approaches in order to allow the holdings of the cultural heritage sector as a whole to be discoverable whatever the community to which an individual researcher belongs.

Approaches to interoperability

Interoperability is best defined as the ability to exchange and use metadata (and data) without manipulation [2, p. 369] or any human intervention beyond that required to create it [3]. It is somewhat more difficult to achieve than interchange, that is the exchange or transmission of metadata which requires processing or conversion to render it usable to the recipient system [4]. To achieve interoperability requires some degree of agreement on semantics between systems and some negotiation between them to ensure that meaning is shared before metadata is exhanged [5, p. 61].

The three metadata schemes described above achieve some, although varying, degree of interoperability within their respective domains by virtue of shared semantics. The MARC standard achieves this most effectively as its semantics are the most clearly and unambiguously defined of the three. EAD has proved itself rather more problematic owing to its origins as an machine-readable derivitive of document- rather than data-centric finding aids [6, p. 111] and the multiple ways in which it allows the same concept to be encoded [1, p. 123]. CIDOC-CRM has achieved some success, noted above, in allowing the cross-searching of museum holding by virtue of the enhanced interoperability that it facilitates by its extensive semantic modelling.

Achieving interoperability across these three domains is more difficult than doing so within each. Varying semantics are one part of the problem: the definition of a title, for instance, is subtly different within MARC, EAD and CIDOC-CRM and so complete semantic interoperability is difficult to achieve. Just as problematic is syntactical interoperability, achieving an agreement on the way in which these semantics are to be encoded. MARC uses a flat-file approach similar to a traditional catalogue card. EAD employs XML with extensive internal hierarchical structures that mirror the traditional stratification of components in a archive, and CIDOC-CRM works as a conceptual model at a higher level of abstraction. All of this establishes what are often called ‘silos’ of data or metadata which can only communicate with each other in a limited way.

One popular approach that is often cited as a mean to break down these silos is to employ the techniques of the Semantic Web. Since it was first proposed by Tim Berners-Lee as an extension of the World Wide Web which would incorporate truly semantic linkages and potentially allow the Internet to function as a single repository of structured data [7], it has often been cited as an ideal medium for breaking down the metadata silos that these three communities have established, particularly to facilitate federated searching [8].

To achieve this involves converting metadata to RDF (Resource Description Framework), a series of tri-partite semantic statements which function as the basic information units on which the Semantic Web is built. These ‘triples’ emulate the structure of a simple sentence in which a subject is linked to an object by a semantic predicate: for instance ‘London’ – ‘is the capital of’ – ‘United Kingdom’. Each component of a triple is usually represented by a URI (Uniform Resource Identifier), a unique identifier which defines its content unambiguously wherever it is found on the Internet.

The mass of ‘triples’ that comprise the Semantic Web can potentially be treated as a single reservoir of structured metadata which can be searched as a single entity, effectively blurring or eliminating completely the boundaries between the repositories in which it is held. It also has the benefit of great flexibility because of its disaggregation of metadata into small units which can be reused and recombined as required. For these reasons, it has been strongly advocated within the digital asset management community and underlies such widely-used repository systems as Fedora Commons [9].

There are, however, significant drawbacks to employing RDF-based models as the basis of metadata architectures. The blurred boundaries of repositories of RDF triples present problems for digital preservation where most practices, including the widely-employed OAIS standard [10], are based on discrete packages of data and metadata with clearly-defined edges. Similarly, this blurring can cause problems for the definition and protection of intellectual property rights as it become difficult to determine the extent of ownership of a collection of triples [11, p. 92]. Considerable doubts have also be raised within the library community and others as to the practicalities of maintaing large and amorphous collections of triples [12].

An alternative approach to interoperability

For the reasons cited above, an alternative approach to breaking or at least blurring the boundaries between the metadata silos of these three communities is proposed in this paper. The work detailed here describes a metadata strategy devised at the Warburg Institute in London for the enhancing the interoperability of the extensive database of iconography maintained by its Photographic Collection.

The Photographic Collection owes its existence to the art historian Rudolf Wittkower (1901-1971) who established, in the words of the Institute's website a "physical photographs of sculptures, paintings, drawings, prints, tapestries and other forms of imagery... [including] tens of thousands of late nineteenth and early twentieth-century photographs and slides, together with hundreds of thousands of images added since the Institute came to London in 1933" [13]. This collection is unique for one of its size in being organised by iconographic subject instead of by the standard arrangement by period or artist.

At the core of the collection is an extensive taxonomy of approximately 18,000 terms which was initially devised by Wittkower and has been in continuous development ever since. This taxonomy underlies the digital version of the collection which currently numbers 80,000 of the 400,000 items in the physical collection. Wittkower's original taxonomy has been expanded to an extensive facetted classification, often extending to eight taxonomic levels in order to allow descriptions at very fine granularities. This taxonomy is much more detailed than the widely-used *IconClass* scheme for iconographic subjects [14] and undergoes continuing extension as new research reveals the need for new facets.

At present the database and taxonomy are contained in a series of mySql tables and accessed by in-house PHP scripts. This has drawbacks in terms of interoperability as the data and metadata within it cannot be readily transferred to other systems or shared outside the Warburg. In addition, the structure of mySql is clumsy at reflecting the hierarchical structure of the taxonomy which, as a result, cannot readily be employed by others when it is encoded in a complex set of relational database tables. For these reasons, it was decided to translate the database's metadata to a more interoperable format in the hope that it, and its taxonomy, may become more accessible in a way which is merited by the extensive scholarly work that underlies them both.

The move to interoperability: CIDOC-CRM

The first stage of the database's move to more interoperable metadata was to design at an abstract level an overall conceptual model for its components and their interrelationships. Rather than doing this by starting with a blank sheet, it was decided to base the design of this model on the museum standard CIDOC-CRM. This would ensure that the knowledge and experience of the practitioners who developed the standard could be utilized to produce a coherent overall model and also, perhaps more importantly, that it would be interoperable at this abstract level with others based on the standard. It would, therefore, act as something of a bridge to the museum community and their metadata practices.

The CIDOC-CRM compliant conceptual model was compiled by Dr Rembrandt Duits, the Deputy Curator of the Photographic Collection and the designer of the database. The result of his extensive work on this is expressed diagrammatically as follows:

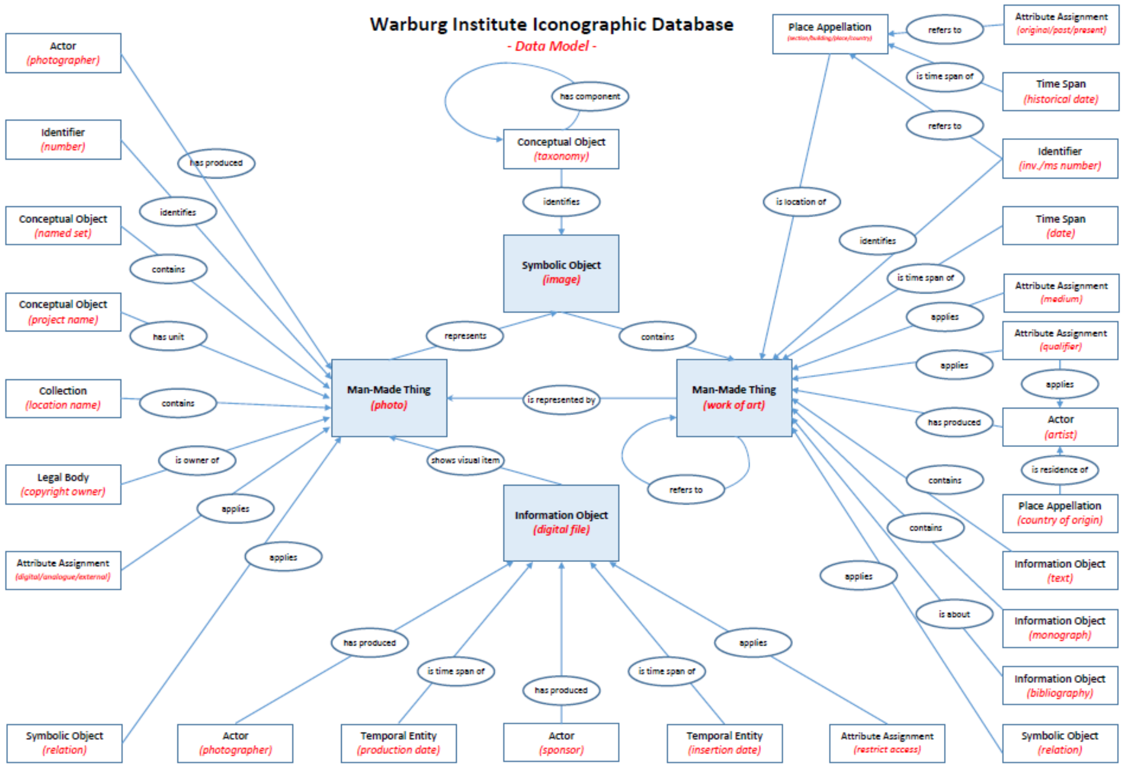

Figure 1Warburg Iconographic Database data model

In this diagram the name of each component of the model is shown within brackets in red and its counterpart class in CIDOC-CRM in black.

At the centre of the model are the four central components of the database, each of which needs to be distinguished as a discrete entity owing to the diverse set of elements to which it must be related. These four components are the image as a symbolic object (to which the classification facets of the taxonomy are linked), the work of art depicted in the image (to which such information as the artist and date of creation are linked), the photo of the work of art (which requires information on its capture) and the digital object in which the photo is stored on the system (with its own array of technical metadata).

Despite the inevitable complexity of this model, every component readily fits into the CIDOC-CRM scheme and finds a place within it. Employing the scheme effectively clarifies conceptual distinctions that might otherwise be blurred, such as between the symbolic object that is an image and its manifestation as a work of art. It also ensures some degree of congruence at this relatively abstract level with other CIDOC-CRM models, so making any mapping of its constituent metadata to others who apply the same scheme easier to achieve.

The move to interoperability: MADS, MODS, METS, PREMIS, METS Rights

Once this model was defined, the next stage was to translate its abstractions into a format that could be used in live systems; this was done by serializing it into an interoperable metadata standard. The chosen format for this was XML, one of the most widely used for metadata as it is software-independent and recognised as one of the archivally robust available. Employing XML ensures that the database's metadata will be readily accessible whether or not the systems currently employed to house it become obsolete in the future.

After choosing XML as the encoding format, the next issue to be faced was the choice of metadata schemas into which the conceptual model could be serialized. It was decided to employ a number of these which owe their provenance to the library sector, particularly that part concerned with the development and maintenance of digital libraries. Two schemas devised by the Library of Congress and hence fully congruent with established bibliographic standards were chosen to encode the Warburg taxonomy and the database's digital object metadata respectively.

The first of these is MADS (Metadata Authority Description Schema) [15], a schema designed to record authority records (such as personal or corporate names) and classification schemes. MADS allows hierarchies of any depth to be encoded, making it ideal for the extensive Warburg taxonomy. This example demonstrates how a second-level topic, Astronomy and Astrology, is recorded with its link to its immediate parent term, Magic and Science:-

```
<mads>

   <authority ID="vpc-cat2-71">

      <topic valueURI="http://warburg.sas.ac.uk/vpc/id/cat2/71"


 authorityURI="http://warburg.sas.ac.uk/vpc">Astronomy and

 astrology</topic>

   </authority>

   <related type="broader"

            xlink:href="#vpc-cat1-9">

            <topic>Magic and Science</topic>

   </related>

</mads>
```

Each term receives an ID (here 'vpc-cat2-71') by which it is referenced from its child terms: the link to its own parent term is achieved by the xlink:href attribute which contains the ID of the latter (here 'vpc-cat1-9). Each term also receives a URI (Uniform Resource Identifier) – here 'http://warburg.sas.ac.uk/vpc/id/cat2/71', a unique identifier by which it may be referenced from anywhere on the Internet.

The entire Warburg taxonomy may be recorded in a single MADS file in this way. Because it is encoded in XML and in this logical and simple hierarchical structure, it is readily transferrable to other systems. By making the taxonomy available in this way, it is hoped that others may be able to take advantage of the extensive scholarly work that it represents.

The second schema employed, MODS (Metadata Object Description Schema) [16], records the bulk of the metadata associated with each image. This schema is derived from MARC21, the primary bibliographic standard used in the library sector, and so is designed to record detailed descriptive metadata. Information recorded here includes artists' names, dates of creation and physical locations of works of art and also details of books or manuscripts in which an image is to be found. It also includes, most importantly, details of iconographic subjects by referencing taxonomic terms in the MADS file using their URIs:-

```
    <mods:subject                                           valueURI="
http://warburg.sas.ac.uk/vpc/id/cat2/71"/>.
```

As important as the content of the metadata encoded in MADS and MODS is the overall structure within which it is embedded: it is this structure that encapsulates the set of linkages within the data model and particularly its differentiation between the four components (image as symbolic object, work of art, photo and digital object) which form its core. This is achieved by embedding the MODS metadata within the framework of another XML schema, the 'packaging' standard METS (Metadata Encoding and Transmission Standard).

METS is designed to provide an overall framework within which all of the metadata (descriptive, administrative, technical and structural) for a complex digital object can be embedded in a logical and consistent structure. METS makes a clear distinction between descriptive metadata describing an information object (the digitized image) and the source from which the digital surrogate is made: the former encompasses the image as a symbolic object, the latter the work of art from which it is derived and the photo which has been converted to digital form.

Separate MODS files are embedded within the METS structure to record the metadata for the image as a symbolic object, the work of art and the photo. Metadata for the digital object, including information relating to its capture and intellectual property rights, is recorded in a section within METS for administrative metadata; two additional schemas, PREMIS (PREservation Metadata: Implementation Strategies) [17] and METS Rights [18] are used for this.

In this way, all of the metadata required by the data model can be encoded within an XML architecture using established schemas from the library (MADS, MODS, METS and METS Rights) and digital preservation communities (PREMIS). Because the underlying data model is designed to be congruent with CIDOC-CRM, the underlying structures of this implementation are consistent with metadata practices within the museum community to which that standard owes its provenance. Its serialization into XML using established standards from the library community also ensures that it is consistent with practices in that sector and that the metadata contained within these XML files can interoperate with that which is held in, for instance, MARC-based library catalogues. In this way, materials that are separated within and between sectors can be brought together and hosted in the same system.

The move to interoperability: EAD

As noted earlier, EAD, the prevailing standard for archival metadata, owes its origins and overall framework to the traditional (printed) archival finding aid. This is particularly manifest in its document-centric approach, which results in large sections of an EAD document consisting of prose descriptions, and its emphasis on the hierarchical division of archival holdings into their traditional levels from *fonds* to series to item to file. It is, therefore, very different in its overall approach and structure from the abstract CIDOC-CRM model and the single-level item or object descriptions in the METS/MODS methodology.

Nonetheless, significant work has been undertaken to establish semantic mappings between EAD and both CIDOC-CRM and MODS which allow a degree of interoperability between these three standards.

An early (2001) proof-of-concept paper demonstrated a considerable degree of semantic overlap between core components of EAD and CIDOC-CRM and demonstrated that the hierarchical description of archival contents in the former could readily be translated into the overall framework of the latter [19]. A later paper from 2011 extended this work to establish a more detailed mapping of the hierarchies of EAD (specifically those of physical, information and linguistic objects) to CIDOC-CRM [20]. These and other mappings have established clearly that semantic crosswalks can readily be drawn between these standards, so bridging their respective communities.

Similar work on crosswalking has also shown that mappings between EAD and MODS may readily be drawn, despite their differing orientations of collection- and object-based descriptions respectively. An article from 2009 designed a successful crosswalk between the two by mapping EAD elements and attributes to their semantic equivalents in MODS, mapping the hierarchical structures of EAD to MODS and emulating in MODS EAD's notion of inheritance of properties from the higher to lower levels of these hierarchies [21]. Using such a crosswalk enables some degree of interoperability between the descriptive metadata recorded within the MODS files in the Warburg scheme and this community standard from the archival community.

One further approach to facilitate this interoperability could be to employ an intermediary XML schema [22], one designed specifically to mediate to an established schema such as EAD. Such schemas offer a constrained, data-centric element set which is designed to be more interoperable than those to which they mediate, but which also can generate metadata conforming to the established schema by the standard technique of XSLT (eXtensible Stylesheet Language – Transformations) transformations. One schema of this type which is specifically designed to mediate to EAD is the CENDARI Collection Schema (CCS) [23]:this could used and possibly extended to fulfil this function in the context of the Warburg metadata scheme.

Although it is evident that further work is required to ensure full interoperability with EAD, these crosswalks and the potential use of intermediary schemas offer ways in which metadata from the archival community may also interoperate with that of its library and museum counterparts. Although the Warburg data model is not based explicitly on EAD as it is on CIDOC-CRM and MADS/MODS/METS, it is nonetheless far from isolated from it: this should enable it to interact with the rich metadata and resources produced and curated by the archival community.

Conclusions

The Warburg metadata model described here, in both its abstract form based on CIDOC-CRM and its serialization into the metadata schemas discussed above, is fundamentally based on the need to move the rich scholarly work underlying the Institute's Iconographic Database from a standalone mySql application to an interoperable form which would allow it to be shared throughout the academic community. To do so, it has employed key standards which are embedded within their respective communities and which are all the fruits of many years' work by their expert practitioners.

It is by the combination of these standards that the full richness of the metadata stored in the Iconographic Database can be maintained in its move to an interoperable form. This is undoubtedly a complex model, one which reflects the complexity of iconography itself and the intricate web of metadata that is required to encapsulate it. But in using established standards, a solid framework is readily built up in which all components find a logical place in a clear, unambiguous structure that can readily be interpreted and understood by other practitioners. Using such standards also ensures the longevity of this metadata and enables confidence that it will be usable long after any current system that maintains and delivers it is rendered obsolete.

It is by the use of such interoperable standards that the boundaries between the museum, libraries and archival communities can at least be blurred and hopefully rendered invisible. To do so within XML is to avoid many of the difficulties that have been expressed about the use of RDF to achieve the same purpose. The approach described here offers a robust and readily manageable metadata strategy which can potentially ‘break the silos’ between these three communities in the long term.

Acknowledgments

The authors acknowledge the work of Dr Rembrandt Duits, the Deputy Curator of the Photographic Collection at the Warburg Institute, which has been central to definition of the metadata model described here.

References

[1] E. J. Shaw, “Rethinking EAD: balancing flexibility and interoperability,” *New Rev. Inf. Netw.*, vol. 7, no. 1, pp. 117–131, 2001.

[2] A. G. Taylor, *The organization of information*, 2nd ed. Westport: Libraries Unlimited, 2004.

[3] S. Bauman, “Interchange vs interoperability,” in *Balisage: The Markup Conference 2011: Proceedings*, 2011, vol. 7.

[4] D. Schmidt, “Towards an interoperable digital scholarly edition,” *J. Text Encoding Initiat.*, no. 7, Nov. 2014.

[5] K. H. Veltman, “Syntactic and semantic interoperability: new approaches to knowledge and the semantic web,” *New Rev. Inf. Netw.*, vol. 7, no. 1, pp. 159–183, 2001.

[6] E. H. Dow, “Encoded Archival Description as a halfway technology,” *J. Arch. Organ.*, vol. 7, no. 3, pp. 108–115, 2009.

[7] T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web,” *Sci. Am.*, pp. 29–37, Jan. 2001.

[8] L. Goddard and G. Byrne, “The strongest link: Libraries and linked data,” *-Lib Mag.*, vol. 16, no. 11/12, 2010.

[9] Fedora Commons, “The Fedora Content Model Architecture (CMA),” 2007-2002. [Online]. Available: http://fedora-commons.org/documentation/3.0b1/userdocs/digitalobjects/cmda.html. [Accessed: 09-Dec-2011].

[10] Consultative Committee for Space Data Systems, “Reference model for an Open Archival Information System (OAIS),” 2012. [Online]. Available: https://public.ccsds.org/pubs/650x0m2.pdf. [Accessed: 07-Jun-2017].

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

[12] R. Hawtin, M. Hammond, P. Miller, and B. Matthews, “Review of the evidence for the value of the ‘linked data’ approach: final report to JISC,” 2011. [Online]. Available: http://ie-repository.jisc.ac.uk/559/1/JISC_Linked_Data_Review_Oct2011.pdf. [Accessed: 27-Jul-2012].

[13] Warburg Institute, “Warburg Institute: Photographic Collection,” *Warburg Institute: Photographic Collection*, 2018. [Online]. Available: https://warburg.sas.ac.uk/library-collections/photographic-collection.

[14] IconClass, “Outline of the Iconclass system,” *Outline of the Iconclass system*, 2018. [Online]. Available: http://www.iconclass.org/help/outline. [Accessed: 21-Aug-2018].

[15] Library of Congress, “Metadata Authority Description Schema (MADS) - (Library of Congress),” 2011. [Online]. Available: http://www.loc.gov/standards/mads/. [Accessed: 24-Nov-2011].

[16] Library of Congress, “Metadata Object Description Schema: MODS,” 2010. [Online]. Available: http://www.loc.gov/standards/mods/. [Accessed: 28-Jan-2010].

[17] Library of Congress, “PREMIS data dictionary for preservation metata, version 2.0.” Library of Congress, 2008.

[18] N. Hoebelheinrich, “METS Rights Extension Schema.” 2004.

[19] M. Theodoridou and M. Doerr, “Mapping of the encoded archival description DTD element set to the CIDOC CRM,” *FORTH-ICS Tech. Rep.*, vol. 289, 2001.

[20] L. Bountouri and M. Gergatsoulis, “The semantic mapping of archival metadata to the CIDOC CRM ontology,” *J. Arch. Organ.*, vol. 9, no. 3–4, pp. 174–207, 2011.

[21] L. Bountouri and M. Gergatsoulis, “Interoperability between archival and bibliographic metadata: An EAD to MODS crosswalk,” *J. Libr. Metadata*, vol. 9, no. 1–2, pp. 98–133, 2009.

[22] R. Gartner, “Intermediary schemas for complex XML applications: an example from research information management,” *J. Digit. Inf.*, vol. 12, no. 3, 2011.

[23] R. Gartner, “An XML schema for enhancing the semantic interoperability of archival description,” *Arch. Sci.*, vol. 15, no. 3, pp. 295–313, 2015.