Integrating Resource Metadata and Domain Markup in an NSDL Collection

Gregory M. Shreve, Marcia Lei Zeng
Institute for Applied Linguistics, School of Library & Information Science
Kent State University, USA
{gshreve, mzeng}@kent.edu

Abstract

Resource level metadata markup alone cannot describe the rich, granular, associative and recombinant information objects potentially contained in modern digital libraries. Today, powerful mechanisms for content and structure description of documents exists in the form of domain specific markup languages such as MatML and MathML. Mechanisms for integrating resource level markup with domain specific markup documents in these languages are required. In the context of the NSDL GREEN digital library, issues and approaches to markup integration are critically discussed.

Keywords: *Domain-specific markup languages, resource-level metadata, structure decomposition, semantic decomposition*

1. Introduction

Metadata schemas for resource description such as Dublin Core, IMS (Internet Global Learning Consortium) and LOM (IEEE Learning Object Metadata) and domain specific markup languages such as MatML (Materials Markup Language), MathML (Mathematical Markup Language) or CML (Chemical Markup Language) have evolved dramatically during the last five years. Much of this development, however, has been a parallel evolution. There is often no clear indication of whether or how resource level metadata should be integrated most effectively with domain-specific content markup or with structural markup meant to describe the internal architecture of resources.

As one of the National Science Foundation's NSDL (National Science Digital Library) collection projects, the GREEN (Green's Functions Research and Education Enhancement Network) digital library project has created LOM metadata records for each of the resources in its collection. These resources are converted into Dublin Core when they are harvested into the joint NSDL metadata repository. While we have found that LOM is an extremely powerful description scheme, during the course of GREEN collection development we found it necessary to consider using other description mechanisms to complement and support LOM. For example, in the GREEN project mathematical equations used in materials science and other disciplines, were not adequately described by LOM. To enhance resource description for mathematicians, MathML was used to describe domain-relevant mathematical content within the resources (articles, dissertations) being described by LOM.

Similarly, a custom markup language DTD (document type description), referred to as GreenML (Green's Functions Markup Language), was written to establish a tag set for a more detailed description of Green's Functions equations than was allowed by MathML. GreenML, for instance, provides for the description of the implementations of programming language source code of Green's Functions, something MathML does not provide.

As a final example, many of the GREEN DL resources contained significant and relevant descriptions of advanced materials and their properties. LOM's Taxon and Keywords elements proved to be descriptively insufficient, so MatML was used to describe this specialized content in greater and more domain-relevant detail.

This specialized descriptive markup creates "new" objects that have semantic and structural associations with their parents. These new objects, specialized markup documents, are also then cataloged into the collection with a LOM metadata record. The integration of resource level metadata descriptions with specialized semantic or structural markup and markup documents is one of the major research problems emerging in digital library development. Defining, representing and manipulating the complex relationships between multiple resource description mechanisms at different information levels of a library collection promises to be a daunting task, but one that will become ever more necessary as the extensible markup language (XML) allows for the creation and embedding of specialized markup inside of and parallel to existing information resources.

This paper will analyze the interaction of resource-level metadata and domain-specific markup languages in knowledge representation for the GREEN digital library. It will examine, in particular, various approaches to integrating multiple description mechanisms considered or adopted by the GREEN project.

2. Discovery of Resources vs. Discovery within Resources

Digital libraries have been primarily concerned, and rightly so, with description mechanisms at the resource level (e.g., metadata), where a resource is understood as a discrete object within a collection of objects comprising the library.

Researchers have identified different classes (or types) of resource-level descriptive metadata elements according to the their function or purpose, as summarized in the following table (**Table 1**):

Table 1. Classes/types of metadata elements

CIC [1]	Hodge [2]	Gilland-Swetland [3]	Greenburg [4]
descriptive	descriptive	descriptive	discovery
administrative	administrative	administrative	administration
structural	structural	preservation	user authentication

According to Hodge, these classes of metadata elements together have the functions of (1) resource discovery, (2) organizing resources, (3) facilitating interoperability, (4) digital identification, and (5) archiving and preservation. Of these classes, descriptive metadata relates directly to describing the information content of the library's resources. Discovery metadata assists in the identification and retrieval of the resources and includes metadata elements that attempt to represent the topical attributes of a resource: descriptions of domain, field, subject classifications, or important domain-specific access points for discovering the record (keywords, diagnostic vocabulary, thesaurus entries or terms in terminological systems). According to Greenburg's study, out of the 15 elements in Dublin Core, 93% (14) are for discovery. In other schemas related to image processing, elements for discovery function comprised 58% (EAD), 73% (the VRA Core) , and 90% (REACH).

However, two barriers exist regarding the discovery function in current metadata schemas and their applications. First, resource metadata, even if it is attached to or embedded in a resource (such as in the head section of an HTML document) or stored in a separate place (such as in a bibliographic database), is always a "surrogate" for the resource, for the actual information-embodying item, or information container. Second, in most digital library metadata systems, including the Dublin Core and IEEE LOM used in the GREEN DL collection, subject/topic metadata elements are very limited. IEEE LOM, for instance provides only the Taxonpath and Taxon elements and the Keywords element to explicitly describe the contents of a resource. Dublin Core only has the Subject element.

On the other hand, no matter how many topical attributes are represented in a metadata record, the information only leads to the "discovery" of the described resource. For example, if a library user were interested in the corrosion resistance properties of certain alloys, he or she might be guided by the resource description to an appropriate resource, such as the ASM Alloy Digest, which contains over 4,300 data sheets on a wide range of materials including plastics and composites. To further explore the information content of the resource, one would have to rely on the internal information structure of the book, not on the metadata. In many digital libraries, and indeed in document repositories of all kinds, access to document information stops here, with discovery. The user then reads or browses the document if it is small, attempts to use the object's integrated or "natural" access mechanism (tables of contents, indices) if present, or uses brute force tools such as full-text searching. In this era of XML and new custom markup languages "stopping" at the discovered resources seems unnecessary.

In order to enable not only the discovery of resources but also the discovery of information entities within those resources, we must use markup languages and their associated metadata to create formal descriptions representing semantic (content) and internal structural organization.

3. Structural Decomposition vs. Semantic Decomposition

The issue of information discovery within resources directly relates to the issue of resource decomposition. Some objects described by a metadata record may be atomic, with little or no internal structure and may not be decomposable into smaller information units. In these cases, it is sufficient to have discovered the resource, for instance, a particular image. Many other objects, on the other hand, are true information containers, information rich and with both a complex semantic structure and a complex internal "document" organization. These objects are structurally decomposable and semantically decomposable. We can use markup languages to create formal descriptions to represent the content and internal organization of these containers. Markup would be used in order to indicate both the presence of possible new information objects within the resource, but, possibly, also to indicate the organization and hierarchy of discrete objects within the resource, that is, the marked up container contains markup objects that may themselves be containers for yet other markup objects.

It might be useful at this time to draw a clearer distinction between structural decomposition and semantic decomposition (recognizing, of course, that structural relationships may carry meaning and semantic elements may imply structure). As a means to illustrate this point let us consider a document resource from the GREEN digital library, John Berger's "Boundary Element Analysis of Bimaterials Using Anisotropic Elastic Green's Functions." Structural decomposition of the document yields specific structural elements: Abstract, Section Headings (e.g., Section 2 Anisotropic Fundamental Solution), Summary and References. The structural elements just listed are, for the most part, culturally bound and reflect textual conventions in American scientific and technical discourse. There are other, less obvious, structural elements also embedded in Berger's article, including numbered formulae and tables.

Markup of this kind of internal document structure has obvious advantages for a digital library. It exposes the internal organization of resources to search, retrieval, and description. Resource metadata, initially applied only to the containing resource, can now be applied to a greater range of (possibly) usable resources within the container. Thus, using another example from the GREEN collection, Lingyun Pan's dissertation, "Boundary Element Strategies and Discretized Green's Functions: Applications in Composite Materials and Wave Mechanics" has been added to the collection as a composite object. Each of the five chapters of the dissertation has been added to the collection as a discrete resource. The optimal representation of the structural relationships between the five "child" resources and the parent document would involve both document description markup of the original document and corresponding resource descriptions for each of the marked up elements, plus indication of the parent-child relationship. It is clear that the issue of granularity arises here. It is possible to carry structural decomposition down to the level of very small structures. Clearly, the purpose and potential user community of a digital library is relevant, and relates to the desired level and focus of decomposition and markup effort. Some relatively granular markup is likely to be useful for a wide variety of users, as for instance, the identification and markup description of embedded images or equations or the identification of special vocabulary or terminology.

Semantic decomposition is less concerned with discovering structural elements (chapters, sections, tables, figures and other document organizational elements) than with discovering and describing useful content elements. Once again, the best way to illustrate this distinction is with an example from the GREEN collection. John Berger's article contains more than just document structure elements like section headings, it also contains content elements, semantic structures with information value to a particular user community. It might be useful for the user community if these content elements were marked up and retained. Semantic structures are not the abstract "meanings" of sentences or paragraphs, this is the linguist's approach to semantic structure, rather, they are pragmatic/semantic constructs. Semantic structures are identified and marked up because they have relevance to the identified interests of users working in particular domains. Thus, in Berger's article, within Section 5, "Example Problem," there is a brief description of a copper-nickel multilayer material. Composition, fabricating process and other property details are given. This particular information might profitably be marked up with a domain specific markup language such as MatML) and thus exposed to discovery and extraction by an interested user in the materials science community.

The domain specific markup languages chosen to describe content elements would be dependent, once again, on the pragmatics of library use, that is, what the library's user communities expect to do with the resources and what functions they would expect the resources to serve. The tag names, attributes and document type descriptions provided by a markup language directly reflect a domain-specific semantics. Markup languages are the single most important way that explicit domain semantics can be applied directly to natural language and multimedia resources. MatML, for instance, is a direct reflection of what content in journal articles, databases or materials descriptions is of greatest concern to the materials scientist.

4. Item-Level Description vs. Content-Level Description

In the GREEN project, item-level descriptive metadata based on the IEEE LOM (Learning Object Metadata) specification is applied to four main categories of resources: problem descriptions, scientific and technical literature, an array of teaching materials, and data sets. The IEEE-LOM metadata specification (IEEE TSC 484.12) provides a rich set of descriptors and is capable of describing, in addition to the core characteristics of the resource (identifier, title, author, description), a wide range of other characteristics from rights management through educational uses, to technical information. It does not provide significant resources for the description of important structural and semantic information inside the resource.

Extending the definition of library "resource" beyond the boundaries defined by the object in its natural state (book, article, lecture, slide show, video presentation) is an important objective for the digital library community. Accomplishing this objective, however, will involve resolving significant problems involving information discovery within the resource, e.g., how does one effectively analyze, recognize, and mark up important information elements in resources in such a way that it could be effectively accomplished without huge expenditures of human labor? The automatic recognition of content elements, for instance, of interest to a materials science community would involve special purpose parsers, front-loaded, most likely with modules for recognizing diagnostic vocabulary, recognizing certain relationships as expressed in language (is composed of, is a property of), and locating formulas.

As an example of special purpose parsing as a means to decomposition in resource discovery, the GREEN collection has applied term extraction algorithms to collection documents, extracting specialized vocabulary, diagnostic linguistic collocations, and usage contexts. These extracted content elements are used to enhance keyword access to the collection and to build a combined glossary and thesaurus to enhance the educational value of the collection. The extracted terminological elements are described using the Term-Base Exchange (TBX) format based on ISO 12220 (MARTIF).

If new structural and semantic objects are discovered by decomposition using special purpose parsers and described by structural and domain specific content markup, they become new resources for the digital library. Resource level metadata records must be added to allow for

the discovery, access and retrieval of these new objects. However, the identification of new content and structural elements within resources not only exposes the objects to search and retrieval, it also makes it possible to extract those resources and combine and recombine them into new objects, new resources, wholly constructed by recombination (**Figure 1**).

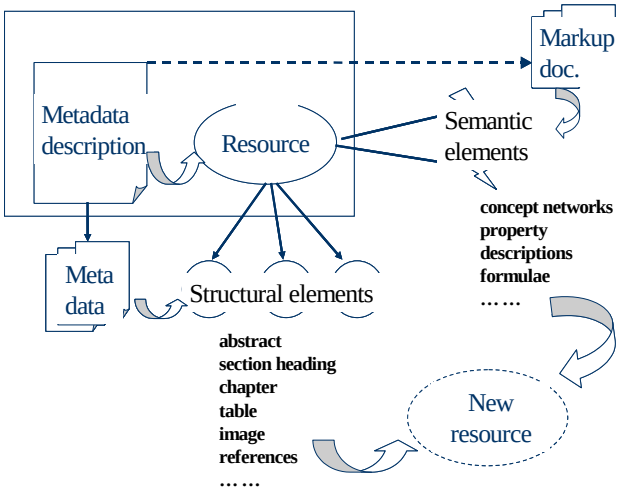

Figure 1. Recombination of discovered resources into new information objects

Ultimately, this means that a collection of digital resources, a digital library, can have a very complex structure with a significant degree of "granularity." At the highest level of granularity would be a set of parent resource descriptions (LOM metadata records). Each of these records points to a specific resource. Any given resource might be a composite object, further decomposable into discrete "child" structural elements (as a book resource could be decomposed into chapters or a video resource into frames or frame sets), or decomposable into discrete semantic objects (concept networks, property descriptions, formulae, images) that might be accessible as resources in their own right. At this point, the digital library becomes not only highly granular, but the new relationships that can be established between objects, either by decomposition or by recombination, make the library highly associative.

When resource level metadata and domain-specific markup languages are applied and integrated appropriately, theoretically they could together offer the following enhanced functions: (a) multiple resource discovery (including discovery of resources and discovery within resources), (b) new resource construction, (c) organization of resources (including internal organization and re-organization), (d) facilitating interoperability, (e) digital identification (including identification of semantic and structural dependencies), and (f) archiving and preservation. This new list of functions expands those offered in Table 1. and are a result a result of the merging of domain specific markup with resource metadata.

5. Integrating Resource Metadata and Content Markup

Objects may be related to one another in multiple ways. These multiple relationships are important to document and exploit. A digital library should provide robust mechanisms to support linkage and navigation between multiply-related resources. In the GREEN project, we have explored several approaches to integrating resource metadata with content markup and revealing the multiple relationships of the discrete cataloged or marked-up elements. Each of the approaches has advantages and disadvantages.

5.1. Approach 1: Extending Metadata Schemas

A number of NSDL projects have already implemented schema extension as a strategy to deal with specialized subjects and contents or the needs of special communities of users. For example, the ADN (ADEPT/DLESE/NASA) item-level metadata schema extends the DLESE-IMS metadata framework, adding geospatial coverage, temporal coverage, and objects in space. GEM (Gateway to Educational Materials) has established an element set fully integrating the Dublin Core qualifier decisions and recommendations of the Dublin Core Education Working Group.

In the GREEN project we extended LOM's nine categories to ten (**Figure 2**). We added a category MATERIALS, which contains elements that are mostly required elements defined in the MatML (Materials Markup Language) DTD. The advantages of using this approach include that the fact that LOM metadata records and MatML markup elements can be processed separately and then later easily linked together. Greater detail about material descriptions contained in the resources are recorded directly in the main resource records, although the resource itself is still marked-up with MatML tags whenever possible. A user accessing an extended record will be able to immediately discover what materials information is available in the record, including specific materials properties or components. This information is available without going into the full text of the resource itself. The "surrogate" provides more detailed information about resource content than the original non-extended LOM record and might help a user decide whether it is worthwhile to access the original text.

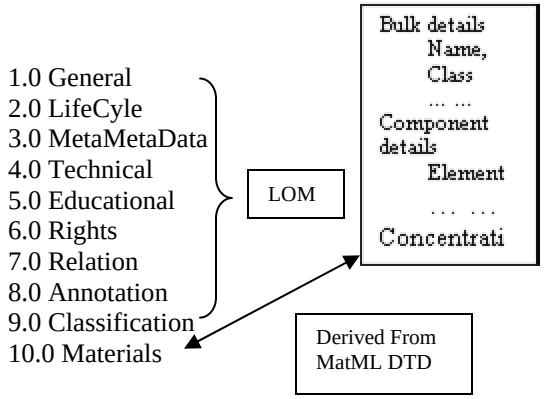

Figure. 2. LOM record extended with a MatML derived category

If desired, the user could easily access the marked-up document and read the sections referred to by the surrogate. If a group of such surrogate records is available, the user could compare material properties across a document corpus. A serious problem with the extension method is that for every category of special interest another category needs to be added to the schema. For small digital libraries with restricted domain scope, and no concerns about expansion of resources, such an approach might be appropriate. In the GREEN project, only one category (MATERIALS) has been added to the original LOM-based schema. However, if more than one markup language, representing another domain, needs to be applied to the collection, then one would need to keep extending the schema.

5.2. Approach 2. Using External Resource Relations to Link to External Markup

In the LOM schema, there is a RELATION category in which elements identify and describe related external resources. This category can be used to reference external markup documents. For example, the LOM resource record for Berger's article contains the following XML (see also **Figure 3**):

```
<Relation>
    <Kind>References</Kind>
    <Resource>
        <Identifier>Green-4-MATML-1.xml</Identifier>
        <Description>MatML description of material
        properties of a copper-nickel multilayer material
        </Description>
    </Resource>
</Relation>
```

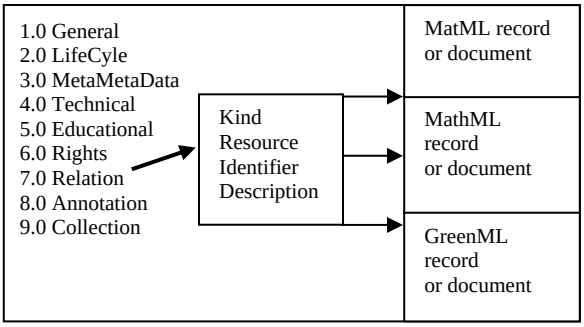

Figure 3. The LOM category Relation references external markup

The value within the <Identifier> tag links the resource (the journal article) with the MatML description. During LOM record display, an Extensible Stylesheet Language stylesheet (XSL) processes the LOM record, generating an HTML page with links allowing the user to navigate from the LOM resource to the MatML resource. The MatML markup (see below) is also processed by an XSL Stylesheet and displayed in a browser in HTML format.

```
<MatML_Doc>
    <Material>
        <BulkDetails>
            <Name>copper-nickel multilayer </Name>
            <Processing>material is fabricated by
            depositing alternating layers of thin-film
            materials such as Cu-Ni, Co-Cr and Fe-
            GaAs</Processing>
        </BulkDetails>
        <ComponentDetails>
            <Name>Cu-Ni </Name>
            <Name>Co-Cr </Name>
            <Name>Fe-GaAs </Name>
        </ComponentDetails>
    </Material>
</MatML_Doc>
```

The advantage of this approach is that open-ended and dynamic markup document linkage can be applied based on an assessment of what contents in the described document should be marked-up. In a GREEN resource, materials descriptions, Green's Functions descriptions, and mathematical equation markup can all be linked to the primary "parent" resource record. The digital library metadata coordinator has great flexibility in applying one or all of the markup languages needed. One possible disadvantage of this approach is that a user must navigate to the linked markup documents to see their contents. In addition, whenever new markup documents are added that relate to a LOM resource, the original "parent" LOM metadata record has to be retrieved in order to add the relevant identifiers.

5.3. Approach 3. Using Other Metadata Schemas To Integrate External Markup

Not long ago, the Library of Congress proposed the Metadata Encoding and Transmission Standard (METS). Each METS record can contain five major sections: descriptive metadata, administrative metadata, file groups, structural map, and behavior section. Of these five sections, file groups and structural maps would be very useful in the integration, recombination or reuse of complex, decomposed resources such as we have discussed. A METS file group section could be used to lists all the document markup files pertaining to the parent digital object. METS structural maps could be used to describe the organizational or hierarchical structure of a resource's associated files (**Figure 4**). For instance, the structural division of the dissertation referred to earlier would be difficult to express using the two previously described approaches.

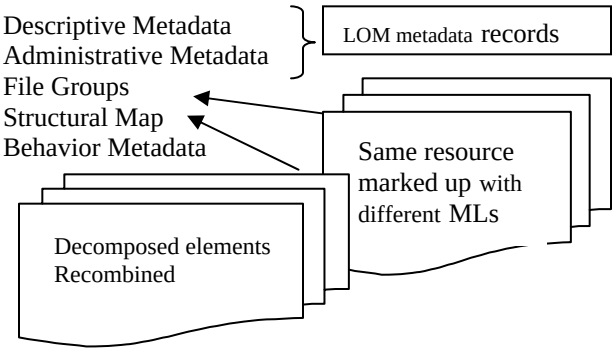

Figure 4. The METS approach can be used to organize a resource's relationships to external files

RDF (Resource Description Framework) linking was also discussed in the GREEN project as a more sophisticated mechanism for representing inter-resource relationships. RDF uses the idea of the XML namespace to effectively allow RDF statements to reference a particular RDF vocabulary or "schema". In another words, when we find some useful elements in different metadata schemas and markup language DTDs or schemas, we could "borrow" and use those elements and form a new metadata format, while indicating where these elements "really" belong by using the XML namespace facility. RDF mechanisms have not been implemented in GREEN, but a METS mechanism is currently being implemented. Both the METS and RDF approach can be used to create documents that indicate all composition, decomposition, and combination and re-combination relations for original or new resources.

6. Issues and Problems

Metadata schemas for resource description and domain specific markup languages have evolved dramatically during the last five years. This has primarily been a parallel rather than a convergent evolution. Even if we manage to effectively decompose complex resources and mark up their contained "child" semantic and structural elements, there is still no clear indication of how resource level metadata can be most effectively integrated with domain-specific content markup or with structural markup. Of course, we can claim that the reasons for developing resource metadata and for doing content markup are very different. Organizing and administering a library or large collection of resources with metadata is different from describing the information content of a single resource. Yet, are these concerns so very different? If the object discovered by a metadata record, as for instance the ASM Alloy Digest, is large and internally complex, composed as it is of 4,300 separate datasheets, then is the designation of the largest "container, "its published form, as the resource to be described, legitimately the stopping point in the digital library description process? One can make the case that the main purposes of metadata as presented in Greenburg's classification (discovery, use, authentication, and administration) could apply equally to information objects within the resource as to the resource itself. Of course, some resources possess a more or less atomic information structure (or at least apparently so) and internal content markup cannot easily be applied (for instance images and video files). In those cases, the resource metadata record, the surrogate, bears the information burden. However even many of those resources generally considered "atomic," as for instance video files, can be both structurally and semantically decomposed using descriptive markup documents. For instance markup using the SMIL specification (Synchronized Multimedia Integration Language) can be used represent the internal structure of video resources with metadata elements representing sequence, scene, shot, frame, and object, actor or person in the frame.

Another problem involves developing robust mechanisms for linking the many disparate objects in the collection into a complex associational web. Linking and navigation mechanisms, and more importantly mechanisms for expressing the semantic nature of the linkages, need to be used to connect metadata external to the resource (about the resource) to the resources themselves, to resources identified within other resources, between resources, and to larger super-structures created from combinations of resources at various levels. Intelligent software agents should be capable of deployment over such a web to carry out information discovery, knowledge creation and document assembly tasks.

The true power that markup languages and metadata description can give to digital libraries will only be realized if structural and semantic markup can be activated as what Tim Berners-Lee has called the "semantic web." Because markup languages begin to allow us to manifest coherent, domain specific semantics within the resource collection, they are a critical element in moving digital libraries closer to Berners-Lee's grand concept. Whether XML and RDF (resource description format) are the technologies up to the

task of creating the semantic web, remains to be seen. However, one thing is certain, we will never achieve the semantic web unless we can radically improve our ability to "find, sort, and classify information" – and metadata markup and linkage is a first step in doing so.

All of these issues mentioned above have led us to explore the inter-relationships between metadata markup and content markup in the GREEN collection, with the objective of achieving a closer integration of the two. Instead of supporting the parallel evolution of resource markup and content markup, we seek convergence. Resource level markup alone cannot create the rich, granular, associative and recombinant collection of resources we should want from a digital library. Using the GREEN collection of mathematical and materials science resources we have been exploring when, where, how, and why IEEE LOM metadata, MatML content markup and document structure markup should be integrated in order to make the digital library more useful for the materials science community.

Sources and References

Sources:

Draft Standard for Learning Object Metadata (LOM). July 2002. IEEE-SA Standard 1484.12.1-2002. Sponsored by IEEE Learning Technology Standards Committee (Available at: http://ltsc.ieee.org/doc/wg12/LOM_1484_12_1_v1_Final_Draft.pdf Last accessed July 21, 2003).

Green-ML (Green's Functions Markup Language) DTD Version 1.0., 2002. Prepared by G.M. Shreve and A. C. Powell (Available at: http://appling.kent.edu/NSDLGreen/GREENMetadata.htm Last accessed July 21, 2003).

METS: Metadata Encoding and Transmission Standard, METS Schema version 1.3, June 2003. Maintained in the Network Development and MARC Standards Office of the Library of Congress. (Available at: http://www.loc.gov/standards/mets/ Last accessed July 21, 2003).

NIST (2001). MatML: eXtensible Markup Language for Materials Property Data. (Available at: http://www.matml.org/ Last accessed July 21, 2003). MatML DTD Version 3.0.,2002. Prepared by E.F. Begley on behalf of the MatML Working Group. (Available at: http://www.matml.org/schema.htm Last accessed July 21, 2003).

References

[1] Committee on Institutional Cooperation (CIC): The Role of Metadata in Enabling Collaborative Digital Library Development. (1999)

[2] Hodge, Gail: Metadata Made Simpler. NISO Press, Bethesda, MD (2001)

[3] Gilliland-Swetland, Anne: Setting the Stage. In: Murtha, Baca (ed.): Introduction to Metadata, Pathway to Digital Information. Getty Information Institute. (2001) http://www.getty.edu/research/institute/standards/intrometadata/2_articles/index.html

[4] Greenburg, Jane: (2001) A Quantitative Categorical Analysis of Metadata Elements in Image-applicable Metadata Schemas. Journal of the American Society for Information Science and Technology 52(11) (2000) 917 – 924