The Information Network Overlay: An Architecture for Contextual Metadata Needed for the Curation Process and Repositories that Support Digital Scholarship

David Gewirtz
Yale University Library
Academic Media & Technology
New Haven, CT
01-203-432-3195
David.Gewirtz@Yale.edu

ABSTRACT

What makes for a successful digital repository is the subject of this workshop. Trust is an essential element because in the absence of trust end users cannot rely upon the authenticity and integrity of repository content for their scholarly purposes. While trust is seen as a necessary prerequisite for a digital repository it is not sufficient for success. This workshop paper suggests that contextual metadata that captures the social context of a digital information object, at the consume phase of its life cycle, is another necessary element for the success of a digital repository. This type of metadata is a dimension of contextual metadata that is also needed to preserve the integrity of a digital information object in the retain phase of its life cycle. In the consume phase contextual metadata provides repository end users with the ability to re-use and process digital information objects for personalized scholarly pursuits. In the retain life cycle phase contextual metadata provides a means to ensure that digital information objects are not corrupted and that end users can trust repository content. Contextual metadata created during the consume phase potentially blurs the distinction between data and metatdata and ironically could make the preservation of digital information objects more complicated. The paper explains and provides a framework to visualize contextual metadata that is based upon the OAIS reference model and research from the National Science Digital Library. In addition the paper briefly addresses three important questions about contextual metadata which are (1) What does contextual metadata look like? (2) How is contextual metadata created? and (3) How can repository users take advantage of contextual metadata? In the workshop use cases that show examples of contextual metadata are presented. .

Categories and Subject Descriptors

H.3.7 **[Information Storage and Retrieval]:** Collection, Standards, Systems Issues, User Issues

General Terms

Design, Experimentation, Standardization

Keywords

Curation, Information Network Overlay, Information Environment, Knowledge, Digital Scholarship, Scholarly Communication.

1. INTRODUCTION: Contextual Metadata

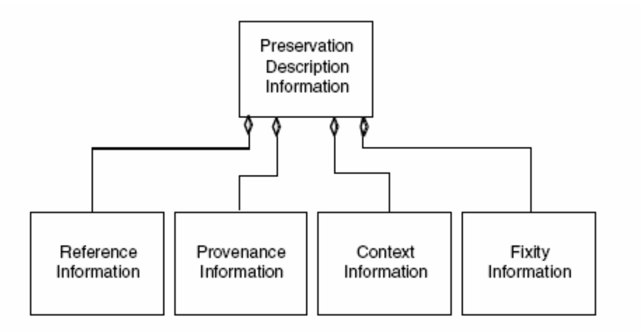
Lord and Macdonald (2003) defined digital curation as the "activity of, managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose and available for discovery and re-use". Said simply, this is the management of data through its scholarly life cycle. Most models of digital life cycles have a chronological and functional component. Chronologically digital objects pass through stages of life that naturally begin with creation, in mid-life they are actively consumed and in the end they are retained for preservation purposes. Functionally at each life stage different stake holders interact with digital objects for different purposes. Contextual metadata in the retain and consume phases (Fyffe etal 2005) of an object's life serves two different but important objectives. In the retain phase contextual metadata according to Waters and Garrett (1996) is a dimension of an object integrity or trustworthiness. In their seminal report on preserving digital information, the authors identified four dimensions of context that needed to be preserved for any digital information object. Informed by the work of (Bearman and Sochats 1995), the ***technical context*** or environment of an object needs to be preserved so that the object can be accessed and rendered to an end user. Contextual metadata for the technical dimension of a digital information object concerns the hardware and software environment that surrounds the object. In the PREMIS data dictionary, it is identified as the environment of a digital object. In a networked based world more and more digital information objects live in HTML documents where one object is often linked to another. These relationships or connection to other objects transform simple objects into complex objects that can only be understood and preserved if their ***linkages*** are also retained along side the object. Similarly the ***communication*** medium upon which digital information is distributed represents important contextual metadata. This contextual metadata is simple to capture when the distribution mode is a CD-ROM and very complex when the distribution mode is the network. Here the capacity of a network's bandwidth and level of security are critical pieces of contextual metadata that needed to be stored with an object.

Interestingly enough the last dimension of context identified in the Waters and Garrett paper as the *wider social environment* is analogous to the consume phase of a digital information object. When an object is released by its creator into a community different stakeholder take different interests in the object. Now the object can be re-used in ways unintended by its creator and be manipulated or changed without detection or authorization. An untoward consequence of consumption then is that re-use and repurposing can lead to loss or destruction of digital information and therefore compromise the cultural record and the pursuit of knowledge (Waters and Garrett 1996). Nonetheless the ability to re-use and process digital information are features that extended it use and significantly add to its scholarly value through the creation of knowledge or learning objects. If contextual metadata is found at different phase of the life cycle is their corresponding difference in its appearance and complexity?

2. What does Contextual Metadata Look Like?

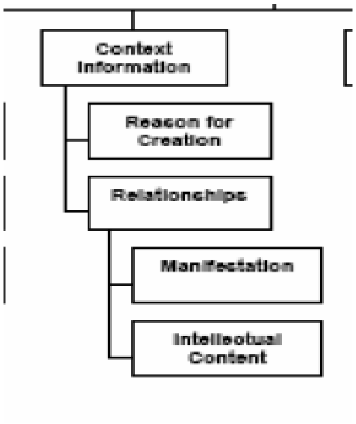
The visualization of contextual metadata is dependent upon the information model that is used to represent this type of data object. The library community can refer to two information models to describe contextual metadata, the OAIS information model and the information model designed by Lagoze et al (2004) for the National Science Digital Library. The OAIS information model consideres contextual metadata as preservation description information which is one of four components of an informaiton object as shown in figure one.[1]

Figure One



The elements of preservation descriptive informaiton, in this model are the same as those elements identified by Waters and Garrett that are needed to preserve the integrity of a digital information object. In OAIS language context information "documents the relationships of the Content Information to its environment. This includes why the Content Information was created and how it relates to other Content Information objects existing elsewhere as seen in figure two." (CCSDS 650.0W5.0 2000). [2]

Figure Two



In this regard contextual metadata is focused upon the retain phase of the digital life cycle. Context in this model is concerned about object relations (structural, dependent or derivative), manifestations of the object and the reason for creation. This is a different sense of contextual metadata that is suggested by the wider social environment introduced by Waters and Garrett (1996). Here the concepts of re-use and processing are present whereas they are absent from the examples of context given in the OAIS view. In contrast the NSDL information model for contextual metadata focuses upon re-use and process or the social life of information. As such their information model of contextual metadata centers about the consume phase of the digital life cycle.

In the NSDL contextual metadata is visualized as an information network overlay (INO) and an information model that has a representational and functional view In non-technical terms an INO is an architecture that allows network resources stored in a repository to be enriched by contextual information. The architecture permits dynamic creation of new repository resource from existing content and provides a community of end–users with the ability to interact with repository content. Here end-users can become agents that can have a role of metadata provider or aggregator. For example, in the role of metadata providers users can attach annotations and reviews to repository content.

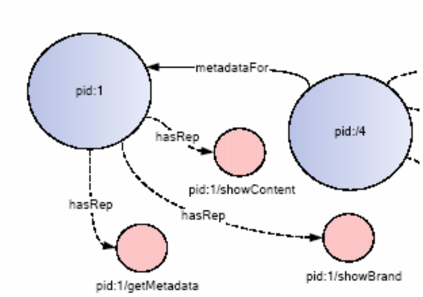
In the informational model the representational view (see figure three (Lagoze etal 2004) uses a directed graph to represent contextual relationships between digital information objects.[3] In the directed graph nodes are really typed Fedora objects and the edges express relationships or context between the objects.

[1] In the OAIS model of an information object the other components are (1) content information or the object source (2) packaging information and (3) descriptive information. These components are combined in packages which serve different repository services. For example a Dissemination Information Package (DIP) is produced by a repository in response to a request for content.

[2] Beyond context, which was discussed in the introduction, the other thre elements of PDI are fixity, reference, and provenance.

[3] A complete explanation of an INO and directed graphs are beyond the scope of this paper. The interested reader is encouraged to read Lagoze etal (2005) where this technology is explained in detail.

Figure Three: Representational View



In the functional view the NSDL content model is used to represent context also through a typed Fedora object. The content model defines typed Fedora digital objects that are configured with a set of data streams, disseminators and relationship ontology.[4] In the NSDL model there are four Fedora content types and a special ontology component that represents eight possible contextual relationships between objects.

Figure Four: Functional View

Persistent Identifier of Fedora Object
FOXML
Relationship Metadata Annotates, assertedBy, augments etc.
Datastream(s)
Disseminator(s)

The first type is a Metadata object which can contain Dublin core metadata about a resource. Resources are the second type of Fedora object. Resources objects are broken down into two sub-types called Agent and Content. A content object (third type of NSDL Fedora object) usually contains a resource such as a text or image file. The fourth type of Fedora object is the agent or a person or organization associated with roles that an agent can perform. An Agent has two sub-types called Aggregator and Metadata Provider. An agent in the role of aggregator forms sets of resources for reasons such as collection management and semantic groupings such as standards and taxonomies. An agent in the role of metadata provider provides metadata and branding information for a resource (Lagoze etal 2005). The NSDL contextual metadata that can be contained in the NSDL content model are: (1) annotates (2) assertedBy (3) augments (4) hasRole (5) metadataFor (6) memberOf (7) providedBy and (8) representedBy. A brief explanation of these inter-objects relationships is given now and than expanded upon in the workshop through use case examples such as an annotation or branding of a resource. The reader should take notice that these relationship provide a means by which contextual metadata can be associated with the typed Fedora objects described above.

Annotates shows a relationship between a content resource and another resource. The annotation is considered the source content and provides a means to comment on a target resource. When a resource is an Agent the **assertedBy** relationship provides a means to link a role to the agent. In the NSDL model this means that the target role is claimed by the source agent. The **augments** relationship links to metadata objects where the source object modifies the target metadata. The relationship **hasRole,** like assertedBy also ties and agent to a role but in this instance the source agent assumes the target role. In the NSDL model the **metadataFor** relationship indicates that the metadata object is metadata for a resource object. Similarly **providedBy** relationship links a source metadata object to an agent in the role of a metadata provider. In the NSDL model agents can also be aggregators and the **representedBy** relationship link an aggregator to a content resource. Aggregators also harvest information (through protocols like OAI-PMH) and the **memberOf** relationship relates a resource to an aggregator (Lagoze etal 2005).

3. Contextual Metadata Creation and Use

The scope of this papers only allows a very high level view of contextual metadata creation and its application. Readers that want to lean more can link to the following references [2,7,10]. This workship paper briefly discusses creation from a conceptional and operational perspective. Conceptually the creation of contextual metadata has similar requirments and challenges as descriptive, technical or administrative metadata but its preservation and sustainablity will be more problematic due to its complexity. Like any other metadata problem knowing the end-user requirment is the first step in the creation process. The sources of end user requirments come from creators and consumers of a repsoitory's informtion. In the OAIS reference model understanding producer and consumer requirments is the responsibiity of the preservation planning function. Once user requirments are understood the contextual information needs to be represented in an ontology. Languages like RDF and OWL can be used to encode these relationships for networked resources. How these relationships are connected to repository content resources is defined by an information model, like the one developed by the NSDL. To create the contextual metadata the ontologies and information model must be translated into an application that is integrated into the workflows of producers and consumers. In this way contextual metadata can be created both by producers and consumers. In the case of the NSDL this capacity allows creators and consumers to interact with content after it has been released into their community.

Researchers in educational technology have identified several approaches to create contextual metadata that attempt to capture the social, instructional and interpersonal context of a learning object. Lagoze etal (2004) summarized these approaches from the literature which include: (1) capturing opinions from teachers or learners about learning objects used to create curriculum (2) recording descriptions of how learning objects are used in instruction can be used to develop lessons plans or lectures (3) identify the community that created the learning object is a means by which content can be branded and (4) provide access to comments and reviews about learning objects. This enables faculty to receive input from students about digital resources used in a lecture or for a homework assignment.

The NSDL is current experimenting with applications that create context around resources that aid in discovery, selection and use of its resources. These applicatons (Lagoze etal 2006) are designed for both creators and consumers of NSDL content. For example, the application Expert Voice uses a collaborative blogging system to create contextual metadata in the form of annotations. Using this system a subject matter expert can create a dynamic presentaton from STEM content while simultaneously creating an annotation for a particular resource. In the end, the blogging system interfaces with the

[4] A technical explanation of the Fedora content models is presented in Lagoze etal (2005).

NSDL library so that both the presentation and the annotation become new resources in their repository. Another application called On ramp enables resources to be created distributively from multiple users and groups in a variety of formats. Like Expert Voice, output from OnRamp, like educational workshop material can based upon NSDL resources. From the perspective of the NSDL this adds additional context to the NSDL resource since it was re-used for a different purpose.

These technological advances in the creation and machine processing of contextual metdata offer the most promise to make this metadata economically sustainable. In addition, the appraoch taken by the NSDL to create contextual metadata generally conforms to a framework suggested by Malaxa and Douglas (2005) that they suggest facilitates the maintenance and verification of contextual metadata as standards change over time. The component of the framework are (1) Flexible metadata schema (2) Metadata schema views, (3) Metadata templates, (4) Collaborative Metadata editing (5) Contextual help and (6) Effective interfaces.

4. Conclusions

This papers set trust as a starting point for a successful digital repository. In the retain phase of a digital information object respositores have an essentail obligation to maintain the integrity and authenticity of digital objects as a means to protect the cultural record from loss and corruption. Ironically while contextual metatdata can broaden the scope and use of repositroy resources it will also make it preservation of digital information objects more complex. This is because contextual metadata systems like an INO can blur the distinction between data and metadata. Consequently how an object is known and understood becomes based upon it context not necessarily its component parts. The challenge to a repository is to find a means to manage and preserve many representation of the same resources to that they are identifiable and known to end users. In addition, the hope that institutional repositories can capture the scholarly outputs of a University that go beyond traditional journal articels requires that repositories can demonstrate trust to their designated community that resources can be preseve for the long term.

In the consume phase of a digtial information object contextual metadata provides as basis upon which repositories servcies can be expanded beyond simple search and discovery to include re-use and processing of resources. An information network overlay is an architecture that supports the storage and creation of services that use contextual metadata. Virtual learning environments like Sakai that are based upon a Service Oreinted Architecture are designed to exploit this capabiity. In additon contectual metadata holds promise to help faculty prepare instructional material such as teaching aids, lectures guides and templates. In this regard academic repostories can distinguish themselves from the commercial aggregators like google and amazon that have become the default search and knowledge service for many scholars. Finally the degree to which contextual metadata can enhance the impact of faculty research and make digital teaching easier for faculty and more effective for student learning is the degree to which repositories will successfully attract end users for content.

5. References

[1] Fyffe, Richard, etal, “Digital Preservation in Action: Toward a Campus-Wide Program”, Educause Center for Applied Research, Research Bulletin Vol. 2005, Issue 19, September 2003.

[2] Lagoze, Carl etal “Representing Contextualized Information in the NSDL; March 2006,Arxiv; http://arxiv.org/ftp/cs/papers/0603/0603024.pdf

[3] Lagoze, Carl etal “An Information Network Overlay Architecture for the NSDL; JCDL ’04, June-11; http://arxiv.org/abs/cs.DL/0501080

[4] Lagoze, Carl etal, “What is a Digital Library Anymore, Anyway? Beyond Search and Access in the NSDL”; D-Lib Magazine, November 2005, Vol 11 No. 11; http://www.dlib.org/dlib/november05/lagoze/11lagoze.html

[5] Lagoze, Carl etal, “An Architecture for Complex Objects and their relationhips; January 2005, Arxiv; http://arxiv.org/abs/cs.DL/0501012

[6] Lord, Philip, McDonald, Alison, “The e-Science Curation Report; http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf

[7] Malaxa,Valentina, Douglas, Ian “A Framework for Metadata Creation Tools”, Interdisciplinary Journal of Knowledge and Learning Objects, Vol. 1 2005, http://ijklo.org/Volume1/v1p151-162Malaxa28.pdf

[8] Powell, Andy, “A Service Oriented View of the JISC Information Environment, November 2005, http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/soa/jisc-ie-soa.pdf

[9] Rusbridge, Chris etal “The Digital Curation Centre: A Vision of Digital Curation” IEEE Computer Society, June 2005; http://www.dcc.ac.uk/docs/DCC_Sardinia_paper_final.pdf

[10] UNC MRC http://ils.unc.edu/mrc/amega_ccrd.htm

[11] Waters, Donald etal “Report of the Task Force On Archiving Of Digital Information” , The Commission on Preservation and Access and the Research Libraries Group (1996) http://www.rlg.org/legacy/ftpd/pub/archtf/final-report.pdf