Linking preservation metadata and collection management policies

Maria Luisa Calanag
Koichi Tabata and
Shigeo Sugimoto

The authors

Maria Luisa Calanag is a PhD student at the Graduate School of Information and Media Studies, University of Library and Information Science, Tsukuba, Ibaraki, Japan.
Koichi Tabata and **Shigeo Sugimoto** are Professors at the Institute of Library and Information Science, University of Tsukuba, Tsukuba, Ibaraki, Japan.

Keywords

Collections management, Digital storage, Library management, Data structures

Abstract

The long-term preservation of digital resources is one of the most important issues facing the library community. In particular, libraries need a preservation strategy for digital objects, since digitization alone provides access but not preservation. The digital library community is also focusing on the problem of designing and implementing long-term archives or repositories. Digital repository management includes the development and enforcement of policies for tasks such as managing access to collection contents and preserving items in the collection. Comprehensive standards and best practices are currently starting to emerge, and ongoing work has deepened the understanding of the needs and requirements that must be met to carry out effective digital preservation. One of these requirements is the creation and maintenance of metadata in support of the preservation process. This paper would like to share findings from earlier and ongoing work that serve as "groundwork" for the current directions leading to the idea of making metadata a more useful and powerful tool to contribute to the technical solution of digital preservation.

Electronic access

The Emerald Research Register for this journal is available at
www.emeraldinsight.com/researchregister

The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0160-4953.htm

Collection Building
Volume 23 · Number 2 · 2004 · pp. 56–63

· ISSN 0160-4953
DOI 10.1108/01604950410514730

Frame of reference

It is apparent that there is no single answer to all digital preservation problems. In fact, the preservation of digital objects involves a variety of challenges, including policy questions, institutional roles and relationships, legal issues, intellectual property rights, and metadata. The policy and legal aspects of digital preservation have been tackled by Calanag *et al.* (2001a) which focused on national libraries as archives since they are generally mandated by law to maintain deposit collections, thereby providing some kind of assurance that these deposited materials will be kept for posterity. Figure 1 shows the key concepts and relationships mainly discussed in the paper. Moreover, it surveyed the status of legal deposit for online publications in several countries. It also argued for the necessity of selection policies for the purpose of preserving web documents, rather than aiming for comprehensiveness citing issues on costs, access provisions, and copyright issues. It found that selection for collection building and preservation is mainly human-driven, and involves the decision-making process for including or excluding electronic material from the deposit collection. There should be a more cost-effective way to do things. This is what motivated the authors to look more closely into collection-level descriptions and proceeded work along these lines.

Calanag *et al.* (2001b) focused on searching for a cost-effective means of ensuring appropriate management of digital resources, coming across collection levels used in the printed world (Table I) which are, in fact, a part of the Research Libraries Group (RLG) Conspectus, a method of describing collection strengths in a standardized manner (Clayton and Gorman, 2002). This should still be useful in its primary role of collection description; however, more information is required for digital collections. By comparing, analyzing, and synthesizing collection level descriptions used by the Berkeley Digital Library SunSITE (UCB), the Arts and Humanities Data Service (AHDS) in the UK, and the National Library of Canada, a new Table that expresses preservation decision and responsibility for the resource at the time of selection has been created (Table II, which we called persistence levels). An example of a policy would enforce that materials in any category except "archived" may be re-designated from one level to another as required to meet changing information needs, remote server accessibility or responsiveness, local resource demands, etc. Hence, materials that receive the "archived" designation cannot be downgraded to a lower status, and in that case, would be the only ones which the library commits to preserve by intending

Figure 1 Key concepts and relationships

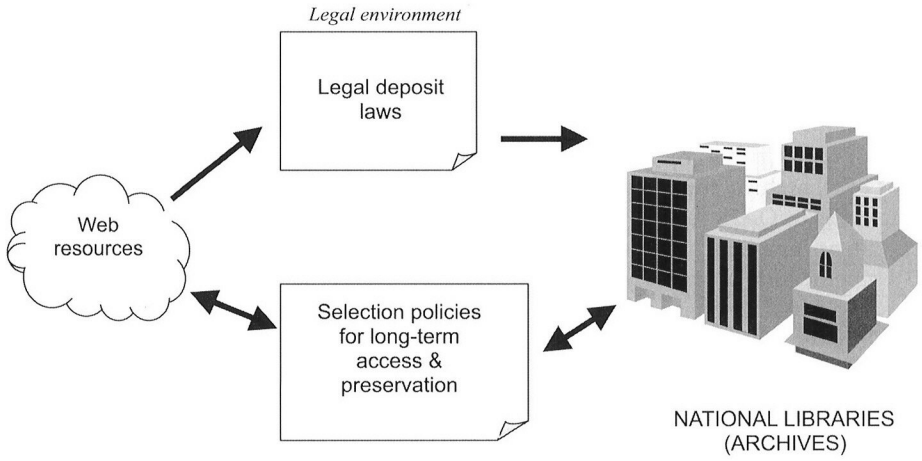


Table I Collection levels

Levels	Description
Comprehensive	A collection to include all significant works of recorded knowledge in all applicable languages for a defined and limited field
Research	A collection which includes the major dissertations and independent research, including materials containing research reporting new findings, scientific experimental results, and other information useful to research
Study	A collection which is adequate to support undergraduate and most graduate course work, and to maintain knowledge of a subject required for limited or general purposes
Basic	A highly selective collection which serves to introduce and define the information available elsewhere
Minimal	A collection in which few selections are made beyond very specific works

Table II Persistence levels

Levels	Description
Archived	Material is hosted in the library, and it intends to keep intellectual content of material available permanently
Served	Material is hosted in the library, but no commitment to keeping it available
Mirrored	Copy of material residing elsewhere is hosted in the library, and it makes no commitment to archiving. Another institution has primary responsibility for content and maintenance
Brokered	Material is physically hosted elsewhere and maintained by another institution, but the library has negotiated access to it; includes metadata and links in the catalog, and library users can locate and cross-search it
Linked	Material is hosted elsewhere, and the library points to it at that location; no control over the material
Finding aids	Electronic finding aids and metadata held by the library to facilitate discovery and searching; this metadata is associated with the library's digital collections or elsewhere, but may be stored, managed and maintained separately from them
De-accessioned	Accessioned resources that have not been retained after review

to keep intellectual content of material available permanently.

A second objective of Calanag *et al.* (2002) was to compare, analyze and synthesize preservation metadata element sets (PMES) of three major projects – the CURL Exemplars in Digital Archiving (Cedars) project, the National Library of Australia's PANDORA project, and the Networked European Deposit Library (NEDLIB) – which all based their metadata frameworks on the open archival information system (OAIS) reference model. The output is a Core PMES which intends to serve as a general metadata framework that can support a broad range of digital preservation activities. The OCLC/Working Group on Preservation Metadata (OCLC, 2002) used the same methodology for a White Paper report.

The comparison showed that the three projects seem to share the view that the primary purpose of preservation metadata is to document the information necessary to facilitate decision-making on the part of preservation managers, and to maintain access to the content of archived

Collection Building
Volume 23 · Number 2 · 2004 · 56–63

digital objects. This is clearly shown by the finding that the three projects focused mainly on the Provenance and Representation Information components of the OAIS information model.

The CEDARS project generally adhered to the OAIS model, but the proposed preservation metadata element set is not intended to include descriptions of all archival functions because there are separate areas in OAIS for the administration and management functions. On the other hand, NEDLIB's scheme focuses strictly on preservation metadata, and not on metadata that have to be preserved. Only the National Library of Australia attempted to develop a metadata set that may be described at collection, object, and sub-object level. This model assumes that the digital object is the primary focus of management and description, and file and collection descriptions are created when appropriate. This finding got the researchers more interested in granularity issues, in particular, collection level descriptions, since they have been widely used in bibliographic descriptions in the traditional library, archive, and museum environments.

This paper synthesized the preservation metadata elements common to the three projects which can be considered "core" or essential for long-term preservation of digital resources. The definitions for our Core PMES are as follows.

Preservation description information

(1) *Reference information*
- Persistent identifier. An identifier or "permanent name" for an object that identifies it uniquely and persistently.
- Date of creation. Date expressed in a standardized form when the manifestation came into being.
- Existing descriptive metadata. Any metadata record which has been generated for the resource.

(2) *Context information*
- Relation. Specifies any other information objects which were judged, at the time of ingest, to be significantly related to the ingested digital object.

(3) *Provenance information*
- Origin. Contains a description of the original digital object prior to ingest; in addition, where the production of the object has involved digitizing, the production process can also be described here.
- Custody history. Contains the identity of individuals or organizations responsible for the storage of the digital object from the date of its creation and the digital archive became responsible for the storage of the digital object, and records when they were responsible.
- Modification history. Describes any changes which anyone responsible for the storage of the digital object made, from the time of creation of the digital object until the digital object became the responsibility of the digital archive.
- Original technical environments. Contains information about the operating environment of the original digital object at the time of ingest, including information on relevant hardware and operating systems, together with the software products that would have been required in order to use it.
- Purpose of preservation. Describes the reasons why the digital object was preserved and deposited in the archive.
- Rights management. Contains links to copyright statement which could include name of publisher, date of publication, place of publication, rights warning, contracts or rights holders, permissions.

(4) *Fixity information*
- Authentication indicator. The mechanism used to ensure the digital object's authenticity.

Content information

(5) *Representation information*
- Object structure. Provides a mechanism for transforming the preserved digital object (stored as a bytestream) into the structured set of digital components needed in order to access and render its content. An example would be information on the object's underlying abstract form description.
- Object semantics. Provides the mechanisms which allow the specific digital object to be rendered. Examples are information on the object's platform, parameters, input format, output format, etc.

Calanag *et al.* (2002) also attempted to provide some structure for the ideas on a general collection management decision guide in the form of a requirements analysis framework that may assist in determining the metadata granularity required for digital resource management within an archive. Interoperability is an important goal. The objective is for metadata and mechanisms to be shared among digital archives, but policies can be tailored to the requirements of the organization.

The next part which has been presented in the International Conference on Dublin Core (2002) gives more details about the guidance model which has been developed with preservation in mind.

Linking preservation metadata to collection management policy

A requirements analysis framework was formulated which associates persistence levels of resources with metadata to help collection managers define the appropriate metadata granularity based on their own preservation requirements. By merging traditional collection levels or conspectus description levels (Table I) and "persistence levels" (Table II) in the form of a matrix (Figure 2), this can serve as a good starting point for developing a method of linking policy to metadata.

In addition, a set of values can be chosen for each combination according to the degree to which digital materials are persistent based on LeFurgy's (2002) definitions. Persistence is based on consistent and transparent rules for description and structure, standardized file formats, and so forth. In general terms, LeFurgy said that degrees of persistence can be represented in three categories. In Table III, these confidence ratings are what we considered as "preservation requirement levels".

Given that persistence is closely tied to the clarity and consistency with standards by digital resources, it follows that materials that are highly structured tend to be inherently easier to preserve and access over time. Conversely, less structured materials tend to be harder to manage (LeFurgy, 2002). In addition, persistence can also be tied to resource availability in terms of the digital object's persistent identifier. The authors proposed that these three preservation requirement levels might determine the granularity of the preservation metadata that will be required to ensure that the digital materials will be preserved and accessed over time. In other words, a choice among *high/medium/low* can be associated with item-, class-, or collection-level preservation metadata, respectively. A general rule of thumb is that as we go from high to low the persistence levels gain lower confidence and stability. Collection manager-defined default ratings and "not applicable" can likewise be assigned according to the institution's policy.

As mentioned, default ratings can be set for certain combinations. For example, if a high rating has been assigned to the combination SERVED + STUDY by the collection manager, this means that Item-level description or metadata should be provided for each resource which would entail a big responsibility and commitment on part of the institution since very detailed metadata has to be provided or created. On the other hand, if it has been decided that LINKED + BASIC = LOW, it means that collection-level description or metadata should be used. UKOLN developed an RSLP Collection Description Tool which can be found in www.ukoln.ac.uk/metadata/rslp/tool/ which can be used to assign descriptive information to collections. These collection-level descriptions or metadata can then be shared among digital archives for cross-searching, access and re-use.

Figure 2 A requirements analysis matrix linking policy and metadata – a sample policy table

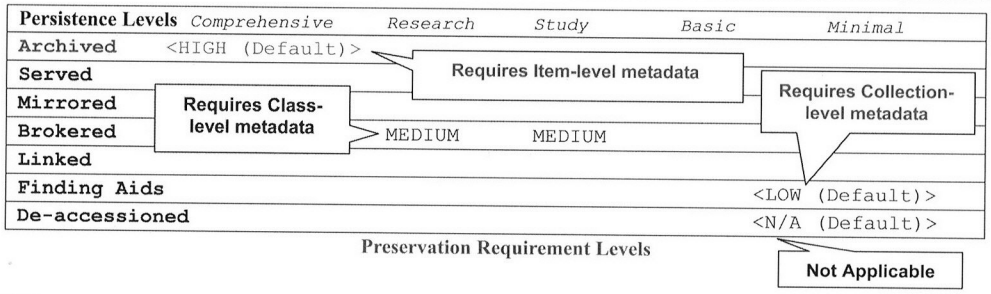


Table III Mapping between preservation requirement levels and metadata granularity

	Metadata granularity	Description
High	Item-level metadata	Individual digital objects are packaged into the content information (CI)
Medium	Class-level metadata	Structural information is handled; this metadata describes types of object attributes, and aggregation information (CI)
Low	Collection-level metadata	Can be added to the descriptive information (DI) and in this paper, this also refers to the RSLP collection description schema

Collection Building
Volume 23 · Number 2 · 2004 · 56–63

Purpose of the study

What had been achieved so far was to lay down a collection management guide in the form of a requirements analysis matrix for general applicability, but where preservation policy decisions can be made according to local requirements. Since interoperability considerations are important in today's distributed architecture, adopting standards are considered to be moving towards this direction.

What we are seeking at present for the next phase is a strategy for ensuring the perpetual access to digital resources that protects the integrity, functionality and meaning of digital materials. This can only occur where acquisition and management of digital resources is controlled, contextual information is secured because sufficient preservation metadata are attached to the resource to ensure that it is capable of interpretation in the future. In other words, we are currently introducing the concept of metadata encapsulated objects. While these are primarily technical requirements, the effectiveness with which these can be addressed depends on the organizational and managerial environment in which they are to be conducted. Preservation strategies without policies will not work.

The next phase would like to test metadata as a mechanism to enforce policies, or to control the behavior of an application. In this study, we continue to view metadata as a special layer in an architectural model of a digital archive system. The purpose is to define a process to enforce different kinds of policy in a uniform way by embedding or encapsulating policy in metadata, wherein the policies and their enforcement processes can be changed easily to accommodate future changes – whether technological or organizational.

To help attain the goal of creating policy-enforcing metadata, the following concepts and models are worth looking into.

Metadata encapsulated objects

Existing work on this would include concepts of the Flexible extensible digital object and repository architecture (Fedora) digital object, the archival information package (AIP) in the open archival information system (OAIS) model, and Michael Nelson's dissertation on "buckets" as smart objects for digital libraries (Nelson, 2000).

Flexible extensible digital object and repository architecture

From the project name itself, the system consists of two fundamental entities:

(1) the underlying Fedora digital object architecture; and
(2) the Fedora repository.

We focus on the first entity. A Fedora digital object is comprised of several components including a unique persistent identifier (PID), one or more disseminators, system metadata, and one or more datastreams. It forms the core of the Fedora architecture, providing a framework that enables the aggregation of both content (i.e. data and metadata) and behaviors (i.e. services) that can also be distributed across multiple platforms via a URI. The Fedora repository provides management and access services for these digital objects. The project has adopted the metadata encoding and transmission standard (METS)[1] as the means to encode and store digital objects as XML entities.

The open archival information system reference model

The most fundamental question concerning preservation metadata is its scope: what types of information are included in this class of metadata, and how is it distinguished from, or overlap with, other classes of metadata. The OAIS reference model has proven to be highly influential in answering this question. It introduces the concept of an AIP, which is the digital object being preserved along with its associated metadata. An AIP includes four separate classes of metadata, two of which collectively embody the informational requirements of preservation metadata. The first, representation information (RI), is metadata that facilitates rendering, understanding and interpretation of the digital object's content. The second, preservation description information (PDI), is metadata necessary to manage preservation of the object, and is the aggregation of four sub-types of information: reference (uniquely identifies the object), provenance (documents the object's history), context (establishes the relationship of the object to other objects and its environment), and fixity (validates the authenticity of the object). In addition to RI and PDI, an AIP includes two other forms of metadata: packaging information, which binds the digital object and its associated metadata into an identifiable package or unit, and DI, which serves as resource discovery metadata in support of the archive's finding aids.

OAIS is a model which is based on the premise that digital objects must be converted into bitstreams which can then be preserved indefinitely. This is achieved by a two-stage process known as "ingest", in which data are separated from medium into an underlying abstract form. The underlying abstract form is then mapped into a bitstream, which is preserved.

Collection Building
Volume 23 · Number 2 · 2004 · 56–63

This model, by operating at a high level of logical abstraction, very successfully describes a system for rendering a digital resource into a format for preservation which can be regenerated by reversing the steps. The portion of the reference model that is of direct relevance to the issue of preservation metadata is the information model embedded within the OAIS framework. The OAIS information model broadly describes the metadata requirements associated with retaining a digital object over the long-term.

Smart "buckets"

A bucket is a storage unit that contains data and metadata, as well as the methods for accessing both. Actual data objects are stored as elements, and elements are grouped together as packages within a bucket. An element can be a "pointer" to another object, or another bucket. By having it "point" to other buckets, buckets can logically contain other buckets. They have the capability of implementing different policies as well. For example, one site might allow authors to modify the buckets after publishing, and another site might have buckets be "frozen" upon publication. Another site might define a portion of the bucket to receive annotations, review, or contributions from the users, while keeping another portion of the bucket frozen, or only changeable by authors or administrators. In short, buckets provide mechanism, not policy.

Policy enforcement

A policy is a set of rules reflecting an overall strategy or objective, affecting the behavior of agents and thus designed to help control and administer a system. A policy rule is a set of actions to be performed by a subject agent on a target agent provided some conditions are satisfied and/or some events are triggered (Harroud *et al.*, 2001). In many business environments, whether working with network management, quality of service (QoS), etc. we often have to face a policy pattern of the kind "When, Who Can/Cannot Do What". As shown in Figure 3, Privileges specify "When", subject specifies "Who", conditions specify "Can/Cannot", and "Do What" is specified by object and actions.

Business and management environments need flexibility and adaptability when implementing their policies. Although the target objects and their structures may be stable, the tasks, including the handling rules and procedures, are varied.

Two methods of enforcing policies are briefly described, as follows.

(1) *Using event-condition-action (ECA) rules.* ECA rules are a technology from active databases and are a natural method for supporting reactive functionality in an XML setting. They can be used for activities such as automatically enforcing document constraints, maintaining repository statistics, and facilitating publish/subscribe applications (Bailey *et al.*, 2002). Instead of implementing reactive functionality directly within a programming language such as Java, ECA rules have a high level, declarative syntax and are thus more easily analyzed. ECA rules automatically perform actions in response to events provided stated conditions hold. Basically, ECA rules on XML databases take the following form:

on event
if condition
do actions

ECA rules have a simple syntax and are automatically invoked in response to events.

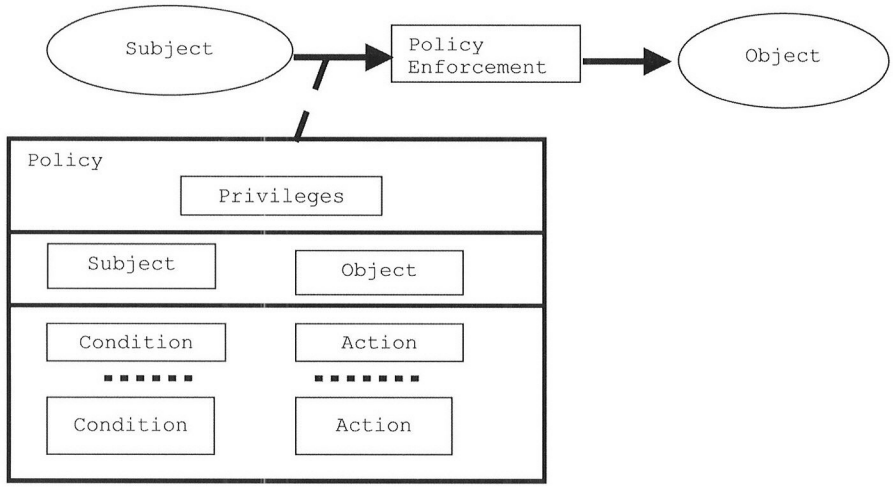
(2) *A policy-driven middleware.* To address the increasingly complex requirements in building and managing distributed applications, next-generation middleware systems are being designed to support policy-driven integration of application-level components with system-level services and resources. Requirements are represented as different aspects of an application. For each aspect, appropriate policy modules are specified or derived from the requirements. From the policy modules, the middleware constructs appropriate policy-enforcement mechanisms, integrating them with application components and system services using meta-level protocols (Tripathi, 2002). Constructing policy-driven middleware involves two notable challenges: creating a specification model to express policies for security and coordination in distributed collaborations, and establishing secure enforcement of the policies.

Related work

Policy carrying, policy-enforcing (PCPE) digital objects

Cornell University is experimenting with an object-centric model of policy enforcement that involves locating policies within the digital objects to which they pertain[2]. Digital objects are able to perform their own policy enforcement by using in-line reference monitors (IRMs). Using the Fedora digital object and repository mode, many different types of digital objects are created and highly customized policies are stored along with base content. Fedora objects not only aggregate data

Figure 3 A policy model



items, but also name code modules that can execute appropriate behaviors for each type of object. The key to this project is getting these bytecode modules to obey the policies that reside with the object. This is achieved by integrating Fedora with Cornell University's PoET software to achieve runtime policy in-lining – that is, the code that "runs" a particular digital object is embedded with checks that prevent violation of the object's policy.

Policy intention architecture (PIA) for digital object repositories

This project, also by Cornell University, is developing a policy architecture that facilitates dynamic policy management and enforcement for large digital library repositories. The PIA provides for dynamic association of policies with digital objects. The goal of this project is to facilitate flexible and dynamic in-line reference monitoring for managed collections of heterogeneous digital objects. Essentially, the PIA allows repository managers, or others, to declare their intentions for sets of digital objects. In the PIA, policies have two components:

(1) a context declaration, and
(2) a set of restrictions to be enforced.

In the context declaration, statements are made about what kinds of objects should be subject to a policy, and the runtime context under which the policy should apply. Restrictions can be fine-grained and tailored to the nature of particular types of objects, and they will ultimately be enforced using in-line reference monitoring on applicable objects. The approach differs from more traditional models in that first, the context is characterized by dynamic object and subject domains; second, restrictions are enforced using IRM; and third, "policy space" is highly modular and typed. Policy enforcement in digital libraries must be scalable, flexible, and extensible – accommodating a wide range of digital objects and usage scenarios.

Virtual encapsulation using XML-based metadata encoding and transmission standard (METS)

METS uses XML to provide a vocabulary and syntax for identifying the digital pieces that together comprise a digital entity, for specifying the location of these pieces, and for expressing the relationships between these digital pieces. There are five key aspects in building a METS document:

(1) expressing the structure;
(2) linking structure with content;
(3) linking structure with descriptive metadata;
(4) linking structure and content files with administrative metadata; and
(5) linking behaviors with structures.

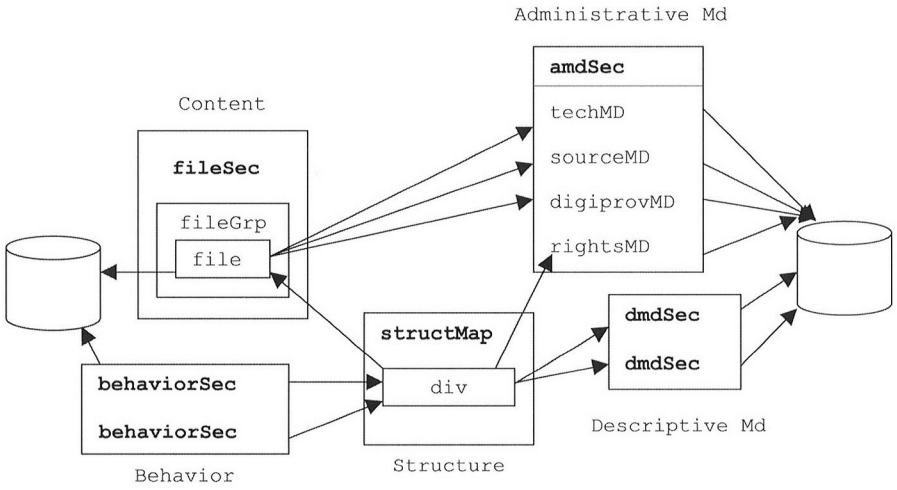
Figure 4 shows the METS framework and diagram.

```
< METS:mets >
   < METS: metsHdr/>      Header
   < METS: dmdSec/>       Descriptive MD
   < METS: amdSec/>       Administrative MD
   < METS: fileSec/>      File list
   < METS: structMap/>    Structural Map
   < METS: behaviorSec/>  Behavior Section
< /METS:mets >
```

ABC ontology and model

Used to model the creation, evolution, and transition of objects over time (time and object transition), "events" could be the key to understanding metadata relationships (Lagoze, 2000).

Figure 4 METS diagram



Conclusions and future work

The authors have laid down a collection management guide in the form of a requirements analysis matrix for general applicability where preservation decisions can be assigned according to local requirements. The concepts and models mentioned in the second half of this paper are currently being worked on as steps forward to enable policy-enforced preservation metadata. Whatever the longer term preservation methods adopted for digital resources, they all need to be wrapped for preservation. Wrapping involves encapsulating or linking the resource to adequate metadata. Another relevant issue in relation to digital preservation is the concept of persistent links. More detailed research needs to be done on this. Libraries, archives, and academic institutions have an interest in the persistence of resource identification, and it has always been thought that metadata will be part of the wrapping or encapsulation of digital objects, and that such objects will be self-documenting.

Notes

1 Metadata Encoding and Transmission Standard (METS), available at: www.loc.gov/standards/mets/, accessed on 4 October 2003.
2 Policy enforcement for complex digital objects, available at: www.cs.cornell.edu/payette/prism/security/peResearch.htm, accessed on 4 October 2003.

References

Bailey, J., Poulovassilis, A. and Wood, P.T. (2002), "An event-condition-action language for XML", *WWW2002*, Honolulu, Hawaii.

Calanag, M.L., Tabata, K. and Sugimoto, S. (2001a), "Digital preservation – some policy and legal issues", *Digital Libraries*, No. 20, University of Library and Information Science, pp. 46-57.

Calanag, M.L., Tabata, K. and Sugimoto, S. (2001b), "A metadata approach to digital preservation", *Proceedings of the International Conference on Dublin Core and Metadata Applications 2001*, National Institute of Informatics, pp. 143-50.

Calanag, M.L., Tabata, K. and Sugimoto, S. (2002), "Linking collection management policy to metadata for preservation – a guidance model to define metadata description levels in digital archives", *Proceedings of the International Conference on Dublin Core and Metadata for e-Communities*, Firenze University Press, pp. 35-43.

Clayton, P. and Gorman, G.E. (2002), "Updating conspectus for a digital age", *Library Collections, Acquisitions and Technical Services*, Vol. 26, pp. 253-8.

Harroud, H. *et al.* (2001), "Policy-based management for multimedia collaborative services", *MMNS 2001*, Springer-Verlag, Berlin, pp. 285-98.

Lagoze, C. (2000), *An Event-Aware Model for Metadata Interoperability*, available at: www.dli2.nsf.gov/ukworkshop/presentations/ (accessed 4 October 2003).

LeFurgy, W. (2002), "Levels of service for digital repositories", *D-Lib Magazine*.

Nelson, M.L. (2000), "Buckets: smart objects for digital libraries", PhD dissertation, Old Dominion University.

OCLC (2002), *Preservation Metadata and the OAIS Information Model: A Metadata Framework to Support the Preservation of Digital Objects*, report by the OCLC/RLG Working Group on Preservation Metadata, OCLC, Dublin, OH.

Tripathi, A. (2002), "Challenges: designing the next-generation middleware systems", *Communications of the ACM*, Vol. 45 No. 6, pp. 39-42.

Further reading

Tripathi, A. *et al.* (2002), "Design of a policy-driven middleware for secure collaborations", *Proceedings of the IEEE International Conference on Distributed Computing Systems* 2-5 July, Vienna, Austria, IEEE Press, Los Alamos, CA.