

Making Metadata: a study of metadata creation for a mixed physical-digital collection

Catherine C. Marshall

Xerox Palo Alto Research Center

3333 Coyote Hill Rd.

Palo Alto, CA 94304, USA

Tel: 1-650-812-4288

E-mail: marshall@parc.xerox.com

ABSTRACT

Metadata is an important way of creating order in emerging distributed digital library collections. This paper presents an analysis of ethnographic data gathered in a university library's educational technology center as the staff develops metadata for a mixed physical-digital collection of visual resources. In particular, the paper explores issues associated with the application of standards, uncertain collection and metadata boundaries, distribution and responsibility, the types of description that arise in practice, and metadata temporality and scope. These issues help to characterize a problem space, and to explore the trade-offs collection maintainers must face when they create metadata for heterogeneous materials.

KEYWORDS: metadata, digital library, ethnographic study, mixed physical-digital collections, visual resources, local knowledge

INTRODUCTION

Metadata is, most generally, data that describes other data to enhance its usefulness. The catalog that emerged as an important component of the modern library is used as a canonical example of metadata, although there are many other well-developed examples within libraries, museums, corporations and other institutions that emphasize intellectual assets as a central part of their stock-in-trade. The development and maintenance of this metadata is, then, a key activity for these institutions. It is the means by which they describe, keep track of, provide access to, and manage their collections.

The Internet (and institutional intranets) has provided additional impetus for developing metadata, as formal and informal document collections and information resources grow unchecked and compete for our attention. How do we

find out about the existence of these distributed collections? How can we find the documents and other resources in them that we seek? Metadata – both human-created and automatically-generated representations and descriptions – acts as a crucial order-maker. Human-created metadata, in particular, is viewed as a way to map the territory of new document genres and new digital media, and invest it with some order that renders it useful and usable [9].

Human-created metadata supplements and goes beyond automatically generated summaries, indices, and other reduced document representations in a variety of different arenas. Resource discovery, use-based retrieval, within-collection organization, retrieval of non-textual media, and interoperability among collections all benefit from the availability of human-created metadata.

First, human-created metadata facilitates resource discovery at the collection or sub-collection level. A collection is likely to be more than an accretion of all it contains; it has been gathered for a purpose. Human-created metadata is thus vital for articulating the scope, intent, and function of a particular collection – attributes that are likely to make the collection easier to locate, and easier to use.

Furthermore, human-created metadata is a natural complement to automated indexing. Materials may then be described in terms of expected use in addition to being characterized by the terms they actually contain. These descriptions can tie the documents together with the particular situation in which the collection has been developed. Fidel explores this distinction between user-oriented and document-oriented indexing in [6]. Besser takes this strategy one step further by contemplating a system that supports user assigned terminology for an even closer link between metadata and use [2].

Human-created metadata is also useful for recording within-collection organization, organization that extends beyond the individual elements, or for noting the relationship between digital and physical realms. In this way, metadata can make a collection easier to browse or aggregate. For example, metadata might be used to express the fact that a particular visual image is part of a series. Hypertext links can be used

to note a variety of relationships among collection elements or to establish a relationship between an electronic record to physical media. Tillet describes seven crucial kinds of bibliographic linkages that can be used to further refine records of within-collection organization [17].

Emerging text, image, audio, or visual image analysis techniques show promise of describing a digital collection element so that it may be classified or found. But these techniques can be greatly enriched by further description via metadata such as keywords or subject-based classification systems. For example, a visual image that represents a painting may be usefully described in terms of its creator – a descriptor out of the reach of current image analysis techniques. Furthermore, subject areas of documents change over time, even as the documents stay the same; human-created metadata reflects the fluidity of classification, even as the document remains fixed.

Finally, and most importantly, human-created metadata furthers interoperability among collections maintained by different organizations [11]. Metadata, for example, may be an essential part of knowledge brokering models; the mapping of one schematic structure into another is a key strategy for managing cross-collection queries. In libraries, the use of MARC records (shared structure) and the Z39.50 store and retrieve protocol is not sufficient to ensure interoperability. Coding conventions and authority lists are a crucial way of ensuring that valuable local practices don't interfere with equally valuable interoperability concerns.

A BRIEF OVERVIEW OF METADATA-RELATED ISSUES

As surely as metadata is valuable, it is also difficult and costly to create. First of all, discovering a workable *structure* is difficult. Even given the well-developed standards in play in library environments, special collections (or special-use collections) may require a choice among several seemingly appropriate standards or a mapping from one standard (say, a standard for describing visual resources like the VRA standard) to another (for example, a USMARC record). Metadata often requires local “tweaks” and adjustments based on the particular collection and its use. Furthermore, constraints are introduced by systems (for example, an OPAC) and practices already in place.

Once a standard is in place, and a metadata structure and strategy has been selected, assigning *values* to catalog the materials in the collection presents the next set of challenges. Assuming that it is not possible or cost-effective to acquire metadata (as, for example, OCLC or RLIN offers), it requires a significant investment to code metadata values into a semi-structured record; external authority sources must be selected, and other local resources that establish a consistent set of values must be marshaled. Some values, for example the subject attributes of an image, pose particular consistency problems. Layne explores these problems in some depth in [8], and suggests that consistency is difficult to achieve for secondary and “subjective” aspects of image subject indexing.

Metadata values are rarely assigned all at once either. A collection grows (and is culled) in an *incremental* fashion, and resource allocation may relegate metadata creation to be a part-time activity or a task that is distributed among several staff members. Once again, this introduces consistency problems.

Furthermore, metadata can and does cross many *boundaries*. Increasingly, collections may consist of both physical and digital elements. They may span genres or media types in a way that stretches existing classification schemes. For example, individual digital images may strain library classification schemes because images represent a much finer document grain-size than books or videotapes.

Metadata can also arise from a variety of *activities*. Traditionally, cataloging is regarded as a principal source of metadata, but it may also be the result of media production (for example, digitization imbues a document with a set of important intrinsic properties, e.g. resolution and format). Less formal kinds of description may result from reading and annotation [13], or use of materials in teaching or presentation [7].

Finally, metadata can take many *forms*. We are accustomed to the structured attribute/value pairs that arise in On-line Public Access Catalogs (OPACs) or in institutional document management systems, but metadata may also usefully include narrative description, such as image captions on a Web page, or informal notes and annotations. It may also arise as implicit organization (for example, the order that slides occur in a slide carousel for a presentation) as well as from explicit coding. As Daniel and Lagoze point out in their justification of Warwick Framework extensions, “...metadata is far too diverse to fit into one useful taxonomy.” [5]

In this paper, I look at the practices associated with creating the metadata for a particular collection of physical slides, coupled with newer digital images, focusing in particular on issues that arise from this heterogeneity. I first describe the study site and the collection itself, and then go on to discuss our observations of metadata creation and its expected uses. From these observations, I draw some conclusions about general directions metadata creation may take, and ways of supporting the practice.

THE STUDY SITE AND METHODOLOGY

How can we open an effective window onto the problems of – and the opportunities afforded by – creating metadata for a mixed digital/physical collection? I have analyzed ethnographic data gathered from a field site at a university library and educational technology center (which I refer to in this paper as “the ETC”) as a means of taking an in-depth look into the practices of creating and using metadata. These data provide details about the overarching institution, a collection of interest, the library and ETC staff’s collection management practices, the collection’s use by the

university's faculty, and the technologies in use at the study site. By way of scene-setting, I describe each briefly.

Methodology. The methodological basis for this study is ethnographic, although the results I cover here are part of a more extensive software co-design project that is organized in a manner described in [3]. The larger participatory effort involves 8 people from a Xerox group, including ethnographers, designers, and computer scientists; I am one of the computer scientists, and as such bring this perspective to the metadata analysis I cover in this paper.

This effort on the library/ETC side involves about the same number of staff members, including the directors of the library and ETC, the head of cataloging, a software specialist, an art and photography reference librarian, and others involved with digital production, information delivery and instructional services.

To understand how the collection is both used and managed as it grows and begins to have digital components, I have used records of interviews with and observations of ETC and library staff members as well as interviews with and observations of faculty members who use the slides in their classrooms and in distance learning situations, including faculty in design, architecture, and biology. Members of our group also attended meetings specifically associated with the staff's efforts to create new digital image resources; these meetings took the form of four kinds of subcommittee meetings covering the staff's work in digitization, cataloging, copyright, and systems. I have also consulted paper and digital documents that are either part of the collection or are used in activities that surround the collection, including the participatory group's email in an effort to bring "net ethnography" (Leigh Star's term) into the picture.

Of special interest to this metadata study were cataloging and systems subcommittee meetings, an observation of a cataloger creating records for an initial set of digital images, and the faculty interviews and observations, since they are significant users of the collection.

The institutional setting. The university library serves a teaching-oriented institution of about 8,000 students. The ETC, which – along with the library – will serve as the principal focus of this discussion, is a service organization that offers photographic, graphic design, media production, and distance learning-related services. The ETC is organizationally separate from, but closely connected to, the library. It is located in the same building as the library.

The collection. The collection has its roots in what started as a circulating slide collection of about 50,000-75,000 35mm slides¹ used by faculty members as visual resources for classroom instruction; this collection is managed by the ETC. Because the university has strong programs in art, architecture, design, and other visually-oriented disciplines, the slide collection is an important resource for the faculty.

The slides (and now, digital images) are either produced or purchased in direct response to faculty needs. To obtain new slides through the ETC, a faculty member fills out a form to request production. As is common practice for teaching collections like this one, images may be photographed from library or archive holdings, in accordance with Fair Use guidelines.

Collection management. The original slide collection was stored in special cabinets, organized according to the Simons Tansey classification system, which was developed especially for slide collections [15]. Because of its nature, the collection is continuously in transition: new slides are produced as requests come in; older slides deteriorate and must be discarded; and additional images are being digitized as part of an extensive internal initiative.

Newer slides are stored in three ring binders, organized according to the faculty member who requested their production. As one would expect, this has resulted in some amount of duplication. A second set of binders, stored in an ETC staff member's office, contains slides that were purchased (rather than photographed from existing materials).

A portion of the collection is managed in the same way as other library materials – through a bar coding scheme that tracks circulation; other portions are tracked using a recording process that involves photocopying the slides themselves. As the group began field work at the library, portions of the collection, and other archival visual materials, were being digitized and stored on servers. These new forms quickly raised questions about the organization of the collection, including cataloging, storage, and retrieval.

Collection use. The collection is used primarily by faculty members and students. Projection for 35mm slides (and, in some cases, for digital images) is available in the classrooms. The faculty members who use slides generally have ready-to-hand ways of putting together presentations – light tables and light boxes – and their own collections of compatible materials (personal slide collections). The use of the visual image resources in distance learning classes is a nascent part of practice; this use is expected to grow.

Metadata technologies. In addition to technologies for dealing with slides *qua* slides, the library and ETC use standard library automation technologies – an OPAC and several other kinds of local databases – for managing the collection and its metadata. The library maintains a Web server to provide access to its OPAC and to mediate access to outside resources. There is a growing trend in the university for departments and individuals to develop their own digital

¹The size estimate given for the collection varies according to the interviewee's organizational perspective. The discussion of collection and metadata boundaries later in this paper reveals that collection extent is not a straightforward assessment to make.

image collections and the means to access them. This trend introduces an interesting set of issues that we will explore throughout this paper – at the very least, it means that the boundaries of the slide/image collection are blurred; its metadata is becoming distributed; and responsibility for the collection is no longer centralized.

As is evident even in this brief overview, to limit this description of the image collection to 35mm slides, or its extent to ETC holdings, is not revealing the whole story. As we discussed the collection and metadata boundaries with faculty and staff at the site, it quickly became clear to us that the move first to library automation, then to a mixed physical/digital collection, confounds any straightforward delineation of the collection's extent as it grows and changes.

Although there are many aspects of the practice of creating metadata that I found compelling, five stand out as consequential for system design in the mixed physical/digital setting described here: (1) the socio-technical constraints on the selection and application of metadata standards; (2) the delineation of collection and metadata boundaries; (3) issues associated with the distributed nature of the collection and the responsibilities for its maintenance; (4) potential types and sources of digital metadata that introduce trade-offs between richness and authority; and (5) the metadata's temporality and scope, given the possibility for much broader (and much more specialized) access.

CHOOSING AND APPLYING METADATA STANDARDS

Discussions of metadata usefully begin by considering standards within the collection maintainers' own communities. Standards not only guide how a collection is described and how individual values are normalized, but also constrain the ability of one institution to interact and interoperate with the collections of similar institutions. For example, the adoption of cataloging standards and a certain amount of metadata centralization has enabled libraries to streamline facilities like interlibrary loan. Similarly, setting up Z39.50 services or making Z39.50 clients available has allowed libraries to query each other's catalogs. Standardization of protocols and metadata sets for digital resources on the Internet is intended to have the same effect – to derive cross-site access benefits from a small number of highly-negotiated universal document attributes.

Hypothetically, an image collection might draw on library standards, since the collection is housed in the library and materials may be accessed and circulated like other library materials. On the other hand, since the materials are visual resources, the collection might well rely on standards that come out of the museum community, which frequently is charged with developing collections of such materials. A collection that includes distributed digital images suggests that World-Wide Web metadata standards, adapted for use with images, might also provide the appropriate model for description.

Indeed, at the university field site, staff members deliberated just this question, and appealed to outside sources in other libraries to help resolve some of the issues. Their deliberations included an evaluation and field-by-field comparison of MESL (Museum Educational Site Licensing) [14], the VRA (Visual Resources Association) Core [18], and the Dublin Core, adapted for image metadata [4]. Hand-in-hand with their evaluation of the descriptive standards, the staff members also consulted a variety of Web-based visual resources and catalogs at other libraries and institutions, including the University of Nebraska at Lincoln's integrated catalog; SPIRO, a Web-based resource containing architectural images; and the SILS Art Image Browser at the University of Michigan.

But is the choice of standards simply an evaluation of what best suits the collection? For our field site, legacy standards and systems that are already in place introduce significant constraints. Furthermore local standards and local practices come into play.

As an important example of such constraints, the staff members needed to come to some accord between their choice of a descriptive standard and the USMARC record, an essential cataloging standard in the modern library [12], and the basis for their OPAC. Their turnkey OPAC represents a substantial investment for the library; as such, it has been the linchpin of a centralized metadata strategy.

Early discussions included proposals for a different metadata substrate than their OPAC – a choice among a growing number of SQL and Z39.50-based image databases, so that the digital images could be stored along with their metadata – and a separate query interface for both the physical and digital components of the image collection. But in the end, compatibility with the existing systems and records tipped the scales in favor of storing the metadata as a VRA-compliant MARC record in their OPAC.²

In addition to the systems and collection-based constraints, the application of metadata standards also must reflect the collection's intended use. The university's library staff identified four types of use-based requirements, including: (1) those associated with library automation and management, including links from digital records to physical media; (2) those stemming from the legal issues surrounding copyright management; (3) those arising from a desire to promote user accessibility; and (4) those that will allow a more extensive digital collection to evolve.

²Shaw argues for the desirability of representing individual visual depictions in the catalog this way but advocates automated mapping of local data into MARC records [16]; Baldonado and her colleagues make architectural provisions for doing such a mapping of distributed multi-standard metadata in the InfoBus architecture [1].

Figure 1 shows how the cataloging standards were put into practice to create metadata records. What is notable here is the combination of physical and digital resources that are brought into play in cataloging. Web resources, the library's existing catalog records, and specific digital files are used in tandem with paper resources like lists, and the physical medium itself.

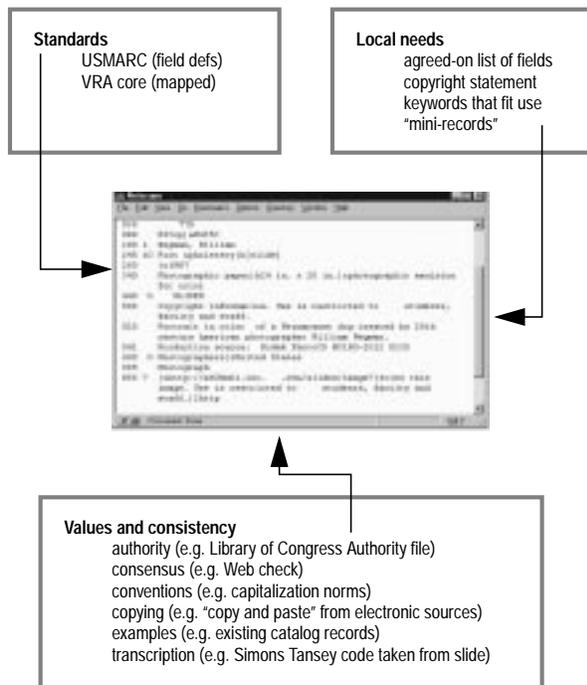


Figure 1. Creating metadata from a constellation of sources: standards, authorities, and local needs.

The staff's discussion of metadata standards underscores the value of continued human mediation between the patron and the collection. First, implicitly in this choice lies the assumption that the cataloger can bring value to the collection through additional description. Second, we also see the acknowledgment of how these descriptions change over time, and how, in fact, some of them may not be captured at all. In the cataloging discussions, staff members have discovered aspects of description that are essentially uncodeable. For instance, they discuss a possible coding that might include the ethnic origin or gender of the artist, so the patron could perform a search that, say, yields women artists of a certain time period – certainly a plausible search in an educational setting. They conclude, however, that these codings are necessarily sensitive to current sensibilities, and must rely on human mediation.

COLLECTION BOUNDARIES, METADATA BOUNDARIES

The creation of metadata relies on the identification of collection boundaries. Naturally, the worth of collection-level metadata depends on the clarity with which it describes the collection's extent; if the collection is described as visual resources for art and architecture instruction, a patron looking for microbiology images is not going to bother to

search for a cross-section of a dicot stem. But also, metadata strategies hinge crucially on the relationship between the individual element and the entire collection. From the initial choice of element descriptors and authority sources to the way in which search tools are implemented, collection and metadata boundaries play an important role.

We found that the collection's transition to a mixed physical/digital resource introduced important ambiguities about its boundaries: Is it a *media-based* collection – the 35mm slide collection or a digital image collection? A *use-based* collection – images a faculty member might use in teaching, say, an Information Design course? A *storage-based* collection – slides and images stored and managed in the ETC's physical cabinets and notebooks and on their own digital servers, as opposed to those stored in faculty offices or on departmental servers? A collection based on *copyright status* – slides and digital images for which copyright issues have been resolved? A *genre-based* collection – slides that covered art, architecture, and design, but not biology? The fieldwork at the site uncovered a number of different interpretations by staff and patrons of what was included in (and omitted from) the collection.

These boundaries are consequential from a metadata viewpoint. The following list enumerates some of the kinds of boundary-dependent metadata that is shifting as the collection begins to include more and more digital materials. This list is organized to first cover three examples of metadata issues connected with collection maintenance and storage: bar codes, production metadata, and metadata that straddles the data/metadata boundary. The remainder of the list provides three examples of complications introduced into access-related descriptors: legacy metadata, catalog records, and transient metadata.

Bar codes. Bar codes and other methods of tracking are necessary for automating the management of physical materials. If the collection is to become a purely *digital* collection, bar coding will no longer be necessary.

Production metadata. The process of digitizing images creates metadata that describes intermediate forms. At the field site, digitization includes the creation of photo CDs by an outside vendor. These photo CDs have printed 'pinnails,' small thumbnails used to index the contents. Thus production itself creates useful metadata; if the production process changes, it will have consequences for the metadata.

Metadata that straddles the metadata/data boundary. Boundary ambiguities between what constitutes the collection and what constitutes metadata for the collection may complicate the design of a metadata store. For example, thumbnails and annotations are variously collection elements and collection metadata. For copyright purposes, it may be useful to see image thumbnails as metadata, in much the same way as an automatically generated summary of a text document is metadata; they are a reduced representation of the materials. However, from a storage perspective, the

thumbnails are just another image resolution, and they are seen in very much the same way as higher resolution images; that is, they are not stored in a metadata store – the catalog – but rather on a collection server.

We might tend to think of annotations as personal metadata, since they elaborate on the content or provide some sort of interpretation and do not generally return to the collection (although there are cases in which marks have been made in books, which are then returned to the shelf). In the digital microbiology collection, however, the annotations have become a valuable part of the content; they are stored as part of the images themselves.

Legacy metadata. Legacy metadata like the Simons Tansey code assigned to a portion of the original collection is of more or less importance, depending on how broadly the collection is construed. If the collection is to grow far beyond its original scope (as opposed to digitizing the current collection), then the old (and difficult to maintain) classification scheme diminishes in importance.

Library cataloging metadata. Since the ETC chose to represent the collection as part of the library's catalog, each image or slide has (or will have) an associated MARC record. Naturally, this metadata spans elements *outside* the collection (including, for example, other library materials like books and videotapes). This choice has ramifications for access – a catalog query that formerly returned six books and a videotape may now also return a large number of image records, a change which may or may not be desirable. The cataloging strategy uses the 440 field to tag individual records as being slide collection records so an informed user can choose to eliminate or include the images in a query.

Transient or situation-specific metadata. Some of the metadata that is desirable for this collection arises from the use of an image in a specific situation. An image of Frank Lloyd Wright's Ennis House may be the 24th slide in a classroom presentation; this sequence information is important for the class presentation, for a subsequent homework assignment, for the final exam, but it may be diminish in value after the end of the term. Transient metadata is discussed in greater detail in a later section of this paper.

DISTRIBUTION AND RELATIONSHIPS

As we move into an era of digital documents, we also move into an era of decentralization and distribution. Formerly, the management of collection items required that the maintainers enforce an idea of a place in which the materials were archived. Even then, in an active, circulating collection materials weren't always literally in the place where they were maintained; only the metadata forms – the catalog records, the circulation records, the shelf lists, and so on – were actually centralized. Given a mixed media digital/physical collection, distribution and decentralization is even more prevalent. Not only may collection elements be stored on different servers or in different databases both within and

outside the managing institution; the collection's metadata can be distributed as well. A single element in a collection may rely upon a composite description, pulled together and assembled as needs dictate; a document name may be stored in one place, its access history in another, and its SGML DTD in a third. From a client workstation, distribution might be made invisible, but collection maintainers can be very much caught up in the challenges this distribution presents.

If we look at the image collection as a distributed collection with both physical and digital components, and we look at the image metadata as intentionally distributed as well, we can begin to realize the full complexity of the situation. Even if the discussion is limited to physical media, to the 35mm art and architecture collection in the basement of the university library, managed by the media center's staff, by design there are still slides in cabinets, slides in notebooks at the front desk, slides in a staff member's office, and slides in circulation. In the broadest interpretation of the collection, it is distributed among a variety of physical and digital stores, some managed by faculty members (for example, personal collections of slides for class, stored ready-to-go in carousels, or digital images on a departmental server), and some by media center and library staff (for example, slides in notebooks, digital images on photo CDs, or digital images on the library's own Web servers).

Metadata, in cases such as this, may be used as the critical 'glue' to bind together physical and digital forms. For example, during the slide cataloging process, a cataloger created records for a digital image that linked the surrogate to three different forms of the image: the *digital* version stored on a library server, the corresponding *physical* slide, and the photo CD's *intermediate digital form*. Thus, the distribution of the materials is among a variety of physical and digital storage places, but the catalog metadata is centralized, serving to represent the relationship among the forms. In a collection such as this one, a way of connecting the various representations to one another is important.

In broader interpretations of collection boundaries, responsibility for the distributed materials becomes an issue. For the ETC staff, responsibility arises from production, acquisition, and possibly copyright status of the materials; for example, if the ETC has produced a slide from a published source at a faculty member's request, they (the ETC) are responsible for the circulation of the slide. That same slide, when seen from a faculty member's perspective, may be part of a personal collection, and outside the purview of the media center; for the faculty member, responsibility for the material is based on a scheme of request and payment. Table 1 shows the many possibilities of where the slides and digital images are stored as a prelude to a discussion of their metadata.

The collection metadata, construed in a similarly (and justifiably) broad manner, is distributed and fluid as well. The desire to maintain a fully cataloged, fully integrated, collection that describes library and media center resources

Form	Location	Responsibility
35 mm slides	cabinets	ETC
35 mm slides	notebooks behind desk	ambiguous (ETC/faculty)
35 mm slides	notebooks in staff member's office	ambiguous (ETC/faculty)
35 mm slides	faculty members' homes & offices	ambiguous (ETC/faculty)
digital images	library server	library/ETC
digital images	departmental server	faculty
digital images	On external server	university-external
digital images	Photo CD	ETC

Table 1: Location of materials: form and responsibility

has led to the early phases of creating catalog records in the library's OPAC. But what happens when the digital collection arises out of distributed efforts?

At the field site, I observed what surely is a common situation. The materials in question are a VHS videotape, stored in the ETC/library facilities and a segment of digital video produced by a university department from that videotape and stored on a departmental server, accessible from the library's Web server. Because the digital video was produced outside of the ETC, there is no link from the catalog record to the digital video; there is, however a link to the location of the physical videotape. Because the digital video is a "one off" effort in a new medium, the departmental metadata store that describes the materials on the departmental server has no record of it. Naturally, the two representations of the video have no connection to each other. The only way for a patron to find the digital video is through traversal.

Thus, in addition to the normal distributed physical metadata that describes the visual materials,³ there is distributed digital metadata, maintained and managed by different groups. Table 2 shows the distribution of possible digital sources of metadata for the collection (assuming broad, but not unreasonable, boundaries). Rather than assuming that the challenge is to unify the metadata records, it seems more realistic to design an architecture like [1] or a Warwick framework-style unification [5] to meet the needs of distributed metadata.

EXTENDED METADATA: THE TYPES AND SOURCES OF METADATA THAT ARISE IN PRACTICE

The activity of creating metadata is not straightforward; there are always collection elements with missing attributes,

³The diverse physical metadata for slides includes, for example, metadata printed on the slide frames, xerographic copies of slides that serve as circulation records, and lists of slides posted in the front of the storage cabinets, and production requests that are stored in the faculty binders.

Form	Location of storage	Responsibility
catalog	library turnkey server	library/ETC
narrative (in html)	library Web server	library/ETC
structured metadata store	departmental Web server	department
narrative (in html)	departmental Web server	department
stand-alone databases	library /ETC PCs	library/ETC
external databases	external servers	university-external

Table 2: Location of metadata: form, control and ownership

descriptive strategies that fall outside of the selected standard, and new ways of accessing and using the collection that stretch the affordances offered by the recorded metadata. No matter how universal a record is – an artwork will surely always have a creator or creators – individual elements often present exceptions. At the field site, for example, some of the design-oriented visual resources in the collection were originally created by an ad agency; whether or not the individual artist should be recorded in addition to the agency is a question that the staff must resolve.

It may be too limiting, therefore, to think of metadata as encompassing only formal attribute descriptions and values assigned according to the recognized authorities. As has been documented in Dublin Core negotiations, metadata most usefully includes a range of description types [19]. The kinds and numbers of fields can be used in a particular way to suit local circumstances, or they may be supplemented by narrative discussion of the image. The field values may not come from a standard source of authority (like the Art and Architecture Thesaurus, or Medical Subject Headings), but rather may arise through local understandings of how the materials are used. Finally, although metadata is regarded as the province of collection maintainers, it may also be derived from the activities of the collection's users.

The metadata being created and gathered at the field site spans a whole spectrum from formal and informal description. Some is being created intentionally – for example, subject classifications – and other metadata comes with the materials, as is the case with the purchased images for Gardner's *Art Through the Ages*. Still more is being generated in the course of normal activities around the collection (like its use in the classroom). The metadata runs from classroom transcripts (created through the collection's use in a distance-learning setting) to narrative (created as part of the process of developing Web-based resources to present the images) to carefully coded attribute-value pairs (created as a result of professional cataloging).

Table 3 summarizes this spectrum. The top rows of the table illustrate examples of metadata that help guarantee interoperability: records that are ostensibly the same across institutions. Subsequent rows show examples of metadata types in which authority becomes more local and more ad-hoc, and while hopes of interoperability diminish, the

Example from image collection	Where the metadata field or type originates	How the metadata value is assigned	Who may assign the value
OCLC MARC records	Purchased record conforming to standards	Values are assigned externally according to standard authority	Vendor
Subject field (MARC 650)	Standards applied in a conventional way	Value is assigned according to authority	Staff
Call number as accession number (MARC 035)	Standards reinterpreted to conform to local requirements	Value is assigned using external conventions and local authority	Staff
Design categories as keywords	Standards reinterpreted to conform to local requirements	Value is assigned using local authority list	Staff (faculty as authority)
Unconstrained keywords	Standards reinterpreted to conform to local requirements	Value is assigned using local knowledge	Staff (and faculty?)
Faculty member storing slide sequence in record	Standard fields used in an ad hoc way	Value is assigned according to highly situated notion of use	Faculty
Narrative description on a Web page	Unfiled metadata	Local knowledge	Faculty
Distance learning chat room transcripts; annotations	Unstructured metadata	Occurs through use in a particular setting	Faculty (and students?)

Table 3: Examples of types and sources of metadata that arise in practice.

possibilities for improved local access and usefulness increase. At both ends of the spectrum are metadata types that remove the burden from the staff: metadata that comes with the materials, and metadata that arises through human activity around the collection

The transition from a physical collection to a mixed physical/digital collection raises some specific issues and opportunities: (1) Metadata that is largely implicit in a physical storage scheme may need to be made explicit in a digital scheme; (2) Metadata values may be more difficult to code, given that increased access to the collection introduces additional use perspectives; and (3) Questions of authority may arise as new metadata types are developed to take advantage of local knowledge. Each issue is discussed below.

Implicitness. Implicit metadata comes from a variety of sources, but two that are particularly noticeable at the field site involve transitions in form. First there is metadata that is derived from the physical and digital storage systems. Necessarily, a good part of the organization that is reflected in the notebooks and cabinets in the ETC may be lost in the transition away from the old classification system to an on-line system. A comparable, equally implicit, ordering scheme now appears in file naming conventions and accession numbers, although it is not clear how this scheme will support ready access. We also observed that some implicit metadata arises as a by-product of the digitization process. The staff assesses the image quality of the images

on the photo CDs, and frequently must adjust various aspects of the image; this transitional metadata is not recorded, although the photo CD itself is stored. As Besser points out, it may be important to record (and indeed standardize) this implicit information about the capture process itself [2].

Perspective. For a collection of this sort – a teaching collection – it is natural for a cataloger to assume a use perspective to assign subject values. But who is the user of the new digital materials? The anticipated improvements in access afforded by a digital collection with digital metadata makes the answer to this question less straightforward than before; the user may be a faculty member, using the collection in a conventional way, or the user may be a student accessing the collection with a different purpose in mind. For example, to describe one of William Wegman’s dog photographs, the cataloger might assign a keyword “dog” and a slide identifier that lets the patron limit the search to the image collection; this assumes the perspective of a patron who is not necessarily interested in the artist *per se*, but rather in locating a picture of a dog. Here Besser’s strategy of user-assigned terminology may be an appropriate solution to broaden perspectives [2].

Authority. Whenever the concept of user-created metadata comes into the foreground, questions of authority accompany it. In contemplating whether or not to open up the metadata creation process to include faculty-supplied keywords, the cataloger weighed the trade-offs between a controlled vocabulary and the access potential introduced by increased descriptors. Even if the keyword-assigner is given a resource like a thesaurus, there remains the question of being able to apply the terms correctly.

As authority-based metadata is supplemented by activity-based metadata, the question of metadata veracity is sure to become increasingly prominent.

THE TEMPORAL ASPECTS OF METADATA

Collections are seldom static; they grow according to the exigencies of use and are culled with media deterioration and as materials are borrowed and lost (or appropriated). Metadata must be similarly fluid. If we see description as guided by use, it becomes clear that metadata requirements will change with shifts in practice. For example, naming authorities may reflect shifts in transliteration patterns. Our study also shows that some types of metadata have limited temporal scope.

Slides (and digital images) are produced in response to teaching requirements and the general needs of the community. The physical forms are weeded as they deteriorate (35mm slides turn red over time, for example); the digital images may change too, as better or higher resolution images become available, or as copyright permissions expire. Metadata can change to reflect changes in the collection, or it may exhibit a different rhythm of fixity and fluidity than the collection itself (see [10] for a comprehensive discussion of fixity and fluidity in

documents). For example, at the field site, catalog records are deleted as slides are tossed; in other cases, metadata is left intact as a valuable surrogate or place-holder even after the materials are discarded.

Some kinds metadata are inherently transient in this use-driven collection – the sequence of digital images used in a distance learning presentation, for example, may only be useful when the course is offered. Similarly, the metadata that describes transitional forms may also be transient. In our study, 35mm slides requested by faculty members are stored in binders according to the requestor’s name. Each binder has its own associated records that arise in during the production process; these records act a little like a shelf list for the binder contents. But the records may have far less intrinsic value after the slides (or their digital counterparts) have been integrated into the larger collection.

Other metadata does have lasting value, but for one reason or another, the metadata element has been abandoned. The problem of what to do with possibly useful legacy metadata is a recurrent one. At our field site, the most prominent form of legacy metadata appears in the form of Simons Tansey coding, the slide classification scheme that had originally been applied to organize the collection. The slide librarian who had the expertise in assigning Simons-Tansey codes has left the ETC, so the newer elements of the collection, including the digital images, do not have this metadata element. Yet the library staff is reluctant to discard the already-assigned values, even though the nuances of the particular classification scheme have become opaque.

METADATA SCOPE

Metadata may not be universal in its scope. Some metadata is local and private, used only by collection maintainers – for example, metadata relating to circulation or to media production – or by a limited set of collection users – for example, the metadata relating to the use of particular materials in a course. Given this specificity, we introduce the complexity of “Who sees what metadata?”

Related to who sees specific metadata elements is how they are used. In particular, are they a fundamental access point for the materials or are they for preservation purposes only? In practical terms, this distinction is realized through indexing – whether a given metadata value is indexed or not.

At the study site, scope discussions centered around whether individual metadata elements would be visible to particular subsets of the slide collections’ user community. Table 4 summarizes the outcome of an initial effort of library staff members to categorize metadata elements as visible or invisible, and indexed or not indexed.

The visible/indexed category is the portion of the metadata that receives the most attention – in this case, fields like the title of the slide and the creator of the artwork. But equally important is the unindexed metadata that simply describes the entity – after all, accession number is the connection

access?	visible	invisible to patrons
indexed	fields for access <i>e.g. title; medium</i>	fields for maintenance <i>e.g. job number; bar code</i>
not indexed	fields for information only <i>e.g. accession number; Simons Tansey code; copyright notice</i>	fields for administration no current examples

Table 4: Organizing the metadata – visibility and indexing

between physical and digital media; copyright notice has already been identified as a vital part of the metadata record; and Simons-Tansey code preserves the legacy classification system. Similarly, the hidden fields – the bar code and the job number – are important to anchor the transitions between the physical objects, which must be tracked and stored, and the digital records that correspond to them.

The distinction between collection managers and collection users forms the first layer of scope: collection maintainers clearly need to see information that the patrons don’t, like the bar-coding that implements circulation. But the ETC saw other distinctions: dividing the world into library patrons and staff is an inadequate definition of scope. Copyright concerns dictate that users outside the university community not be given universal access to the collection itself; so there is already an inside-outside distinction that must be maintained.

However, the inside-outside distinction is not enough either. Images of local architecture, developed by a faculty member, have an associated street address; this might enable a library patron to, say, find the actual building from its surrogate. In practice, this information may be available to a particular set of “inside” patrons (e.g. faculty members), and not others (e.g. students).

In other instances, the faculty-student distinction may not capture metadata scope adequately. If we start considering distance learning classroom transcripts as a possible source of metadata, an even finer-grained distinction is necessary. Metadata visible to class members only then becomes a possibility. What happens, then, when the class is over? Does this metadata become visible to anyone outside the original class participants? Does it expire? Become invisible? Become copyrighted? Any number of complications in scope are introduced by new records of informal metadata.

CONCLUSION

Where does this analysis of ethnographic data leave us? The cost and complexity of creating metadata is, as is generally acknowledged, high. Naturally – as many others have observed – there is no single set of attributes, no one protocol, no clever heuristics that can act as a ‘silver bullet.’

What we can ask is: is there anything that we have learned that *will* help the maintainers of mixed digital-physical collections? An ethnographic approach to understanding the human creation of metadata, and its subsequent use within a

community or institution, is an important way of prescribing the limits of metadata – what is necessary and appropriate for describing the collection for both the maintainers and for the users.

More crucially, ethnographic analysis also provides ways in which we can identify new kinds and sources of metadata that arise out of the collection's development and use in the world. At the field site, naturally-occurring metadata – the distance learning transcripts, the rich local knowledge that comes from use – shows great promise of enhancing the collection's description and access without increasing the already prodigious burden on the collection's maintainers.

What I have highlighted about the field site seems inherent to collections and their metadata – that choosing and applying standards involves a complex balancing act of the collection, its use, and the systems already in place; that collection and metadata boundaries are unclear; that collections and their metadata are growing in a distributed fashion, controlled by many different interests; that descriptions vary widely in type and degree of authority; and that issues related to metadata temporality and scope are going to play an increasingly prominent role in digital collections. In short, there's no easy answer. Instead of focusing on schemas or protocols, this analysis has concentrated on understanding and characterizing the *dimensions* of the problem. Ultimately it is through understanding the entire use situation that metadata can be designed to support the management and access of the complex, heterogeneous collections we expect to encounter in the world.

ACKNOWLEDGMENTS

I would like to thank members of the Work Practice and Codevelopment group at Xerox and the project participants at the field site. I'd especially like to thank Francoise Brun-Cottan, Brinda Dalal, and Pat Wall for on-going conversations about this analysis. I'd also like to thank David Levy for helpful comments on an earlier draft of this paper. Anonymous DL'98 reviews and John Leggett's thoughtful suggestions have been instrumental in the revision process.

REFERENCES

- [1] Baldonado, M., Chang, C., Gravano, L., & Paepcke, A. "Metadata for Digital Libraries: Architecture and Design Rationale." In *DL97 Proc.*, New York: ACM, pp. 47-56.
- [2] Besser, H. "Image Databases: The First Decade, the Present, and the Future." in P. B. Heydorn and B. Sandore (eds.), *Digital Image Access & Retrieval*, Urbana: Univ of Illinois, 1997, pp 11-28.
- [3] Brun-Cottan, F. and Wall, P. "Using Video to Re-Present the User." *CACM*, 38, 5, May 1995, pp. 61-71.
- [4] *CNI/OCLC Metadata Workshop on Metadata for Networked Images*. <http://purl.oclc.org/metadata/image>.
- [5] Daniel, R. Jr. and Lagoze, C. "Extending the Warwick Framework: From Metadata Containers to Active Digital Objects." *D-Lib Magazine*, Nov. 1997. <http://www.dlib.org/dlib/november97/daniel/11daniel.html>.
- [6] Fidel, R. "User-Centered Indexing." *JASIS*, 45, 8, Sept. 1994, pp. 572-576.
- [7] Furuta, R., Shipman, F., Marshall, C., Brenner, D., and Hsieh, H. "Hypertext Paths and the World-Wide Web: Experiences with Walden's Paths." *HT '97 Proc.*, pp. 167-176.
- [8] Layne, S. S. "Some Issues in the Indexing of Images." *JASIS*, 45, 8, Sept. 1994, pp. 583-588.
- [9] Levy, D. "Cataloging in the Digital Order." in *Proc. of DL '95*, Austin, TX, June 11-13, pp. 31-37.
- [10] Levy, D. "Fixed or Fluid: Document Stability and New Media." *Proceedings of ECHT'94*, Edinburgh, Scotland, Sept. 18-23, 1994, pp. 24-31.
- [11] Lynch, C. A. "The Z39.50 Information Retrieval Standard. Part I: A Strategic View of Its Past, Present and Future." *D-Lib Magazine*, April 1997.
- [12] Library of Congress, MARC Standards, <http://lcweb.loc.gov/marc/marc.html>.
- [13] Marshall, C.C. "Annotation: from paper books to the digital library." *Proceedings of DL'97*, Philadelphia, PA, July 23-26, 1997, pp. 131-140.
- [14] The Museum Educational Site Licensing Project. <http://www.gii.getty.edu/mesl/>.
- [15] Simons, W. and Tansey, L. *A Slide Classification System for the Organization and Automatic Indexing of Interdisciplinary Collections of Slides and Pictures*. University of California, Santa Cruz, August, 1970.
- [16] Snow, M. "Visual Depictions and the Use of MARC: A View From the Trenches of Slide Librarianship." *Art Documentation*, Winter, 1989, pp. 186-190.
- [17] Tillett, B. "A Summary of the Treatment of Bibliographic Relationships in Cataloging Rules." *LRTS* 35, 4, pp. 393-405.
- [18] Visual Resources Association Data Standards. <http://www.oberlin.edu/~art/vra/dsc.html>.
- [19] Weibel, S., Iannella, R., and Cathro, W. "The 4th Dublin Core Metadata Workshop Report." *D-Lib Magazine*, June, 1997. <http://www.dlib.org/dlib/june97/weibel/06weibel.html>.