Digital Image Collections:
Cataloging Data Model & Network Access

[preprint version]

Stephen Paul Davis, Columbia University Libraries, 6/95

Published in: *RLG Digital Image Access Project: Proceedings From an RLG Symposium Held March 31 and April 1, 1995, Palo Alto, California.* Edited by Patricia A. McClung. Palo Alto, CA: RLG, 1995. P. 45-59.

INTRODUCTION

Efforts to begin creating the national digital library--a concept that encompasses archival collections, texts, images, sound files, full-motion video, composite documents, and other document types--have led quickly to the need for new approaches to the bibliographic control of such items, at both the local and national level.

The discussions and prototyping carried out during RLG Digital Image Access Project (DIAP) suggest a new data model for the bibliographic control of and access to digital image collections, one based on current, standard cataloging practice but enlarged to incorporate the additional detail, hierarchy, and version information needed to adequately describe digital collections. Though not addressed as part of DIAP, the same model may well be extensible to the cataloging of other types of digital items as well, especially digital reproductions of items created originally in non-digital form.

The proposed data model could readily be "housed" in an SGML-encoded bibliographic (metadata) record that encapsulates both summary bibliographic information along with detailed hierarchical and version-related data, when such data is appropriate and considered useful to record. The record would also include links to the actual digital items, to other related bibliographic records or, in fact, to diverse, related digital objects (such as external electronic publications, databases, numeric files, etc.) The working designation SGML Catalog Record (SCR) is proposed for this new type of record.

The SCR would, by flexibly incorporating data-element "clusters," allow a more narrative approach to the recording and presentation of complex bibliographic information than is practical in the current AACR2/USMARC model, which requires the fragmentation of hierarchically- related components and version information into separate, discrete records. That the current USMARC model serves libraries and users poorly--most notably in the cataloging of microfilm reproductions of printed texts and with complex serial publications--has been widely discussed. Attempts to rectify the situation have been unsuccessful, in large part because of the intrinsically flat structure of USMARC and the automated library systems that have been designed around it.

This paper suggests that, if institutions begin to create SCRs, they would for the present, and perhaps indefinitely, also continue to create summary USMARC records for loading into local library OPACs and national online union catalogs. The USMARC records would always contain pointers to local SCRs, where they exist, rather than pointers to individual electronic texts, images, etc.

Not only would the use of an SGML-based cataloging model better suit the description of digital collection and items, it would allow libraries to better take advantage of the many new non-library-based automated systems, standards and software tools, such as the World Wide Web, now becoming available. Without greater flexibility in the cataloging and encoding of digital documents, library-generated bibliographic data will not be easily integrated into the developing local and national information environment as effective inventories of and indexes to the electronic holdings of libraries.

This proposal is an outgrowth of the RLG Digital Image Access Project, though it should not necessarily be seen as a Project recommendation. It builds conceptually on the Berkeley Finding Aids Project (BFAP) data model and was refined in meetings and discussions with the Columbia DIAP project team (Rob Cartolano, Janet Gertz, Angela Giral, Carol Mandel, David Millman, Janet Parks and

Bob Wolven).

CATALOGING DATA MODEL

Hierarchies. The DIAP participants made several working assumptions about the nature of the cataloging needed for digital image collections. One was that there was no way to anticipate--in system terms--the intellectual structure of a digital image collection or the number of hierarchically-related records that would be needed to describe it fully. For example, an image collection at one institution might be cataloged with a collection level-record and several subcollection or series level records, with links to full records for selected individual items. Faced with the same type of collection, another institution might choose simply to create a collection level record and caption-level records for the individual items.

FIGURE 1: Complex Image Collection, Hierarchical and Flat Record Structures

This unpredictability of the content and structure of archival catalog records, well-understood among archivists and archival catalogers but somewhat less well appreciated by the mainstream of library automation and technical services, took on new significance in a project attempting to integrate cataloging data and linked digital images from multiple institutions into a single database search, retrieval and display system. For example, although the project members initially specified that searching within Visual Photologue be able to be limited to specific hierarchical levels, it became clear that this was not only relatively useless but even confusing to the searcher because of the unpredictability of data content at the same hierarchical level.

Level of detail. Another DIAP assumption was that there was no way to predict the level of cataloging detail that would be provided by an institution at any particular hierarchical level from one collection to another or even within a single collection. The level of cataloging detail would always be subject to institutional practice, cost considerations, the character of particular collection, the importance of the collection within the owning institution s holdings, etc.

Practically, this meant that all defined data elements would need to be repeatable at all levels of the model. Participants agreed that the model implied no prescriptive set of data elements at any level, except as required by national standards, network practice, etc., usually as applied to the highest-level information.

FIGURE 2: Repetition of Data Elements at Different Hierarchical Levels

Although all data elements needed to be available at all levels, in modeling record display within Visual Photologue it nonetheless became clear that displays needed to be sensitive to the presence or absence of data elements at the various hierarchical levels, rather than simply presenting an inflexible template of possible data elements. The SCR proposal below implicitly addresses this issue by assuming that the "narrative structure" of the bibliographic data itself generally reflects the most appropriate display, with relationships and data element "inheritance" implied by proximity and other visual clues.

Multiple versions. The DIAP data model implicitly allowed for description of multiple versions of images, i.e., original images (e.g., slides, photographic prints, maps), intermediate reproductions (e.g., 35mm film), and one or more purely digital versions of an image (e.g., a 600 x 300 dpi JPEG file). Although the project itself produced only a single, low resolution digital image for each image, at least one of the participant institutions (Columbia University) also had items rescanned into the Kodak PhotoCD format, which normally provides five versions of the image at increasingly greater resolution; this resulted in up to six digital versions for each of Columbia's original images.

In addition to description of the original and one or more digital versions, participants also confirmed the desirability of including descriptive information about intermediate images (i.e., slides or film created from originals, from which digital images were to be scanned).

FIGURE 3: Multiple Digital Versions of the Same Image

Data Element Definitions. The participants agreed that, whenever possible, data element definitions would be those described by current standards, including AACR2 and USMARC.

However, it was also agreed that, in light of the other working assumptions, the record boundaries specified or implied by those standards would be ignored for purposes of testing retrieval and display systems.

The lack of standards for the physical and in some cases intellectual description of the digital items themselves was noted by the participants. No systematic attempt to create a data dictionary or a prescribed set of elements was undertaken by the project, in part because of time constraints but also because of the awareness of other efforts to carry out this analysis nationally. The enormity of this task became clear, however, during DIAP-related discussions, where the need for various types of sets additional data elements was noted. The examples below are suggestive rather than inclusive:

CATEGORY	DATA ELEMENTS
digital image physical characteristics	size compression resolution etc.
digital image capture information	date/time responsibility camera type lighting adjustments color correction image editing (cropping, etc.) etc.
digital image subject analysis	image genre content style time period, etc. (Overlap with description of original image.)
intellectual property information	copyright owner(s) permission status terms governing reproduction etc.
intermediate physical characteristics	film type resolution etc.
intermediate capture information	date/time responsibility camera type lighting adjustments color correction etc.

The discussion surrounding the possible need for these data elements inevitably prompted anxiety among some of the participants because of the level of expense and complexity their inclusion might entail. Again, however, it was recognized that some institutions would want and need such detailed information, where others would not. Barring the development of formal national guidelines for the inclusion of such information, all data elements at the item or piece level would have to be considered optional.

Description of originals. Another virtually inevitable set of issues raised by the project concerned the description of the original items from which the digital images were created. Consistent with other variable approaches to cataloging and description, participant institutions had different practices here, sometimes on a collection by collection basis. Participants agreed that, however

desirable it might be to have piece-level description for their image holdings, it was in many cases a practical impossibility.

What did seem clear, however, was that if originals were in fact described at the piece level, this information needed to be recorded within the same hierarchical context as the descriptions of the digital images derived from them. Under the best of circumstances, a full-set of descriptive and subject oriented data elements for each original image would be present in the record cluster. Intermediate, second- and subsequent-generation reproductions of the originals would be coded as hierarchical children of the originals, inheriting the description and access elements of the originals.

While participants did decline to recommend universal piece-level cataloging as part of the project, there was a sense that the creation of digital image access techniques envisioned by this project did have certain implications for the cataloging of originals, however, namely that caption-level cataloging (however brief) was highly desirable to enhance the display of individual digital images. In some cases, however, the captions would necessarily be non-unique (e.g., in collections of closely related images), uninformative (e.g., a sequential ID number), or difficult to interpret without the inheritance of higher-level information. For example:

[IMAGE]
View of Parthenon, 1846

[IMAGE]
NYCG94-1232-12.5

[IMAGE]
View from Southeast

(NB: In the last instance, the collection title is: Empire State Building Photograph Collection; the image group is: Waldorf-Astoria Hotel Deconstruction Photographs)

Cataloging data model. In summary, the cataloging data model developed by DIAP may be described as follows:

- collection cataloging consists of a set of data element clusters
- clusters may be flexibly arranged into multiple hierarchical levels
- a full set of data elements is accommodated, though not required, at each hierarchical level
- descriptions of originals, intermediates, reproductions, versions, and components may be included
- within clusters, AACR2 and related cataloging guidelines and USMARC data element definitions are observed whenever feasible
- relationships between data element clusters may be implicit or explicit
- links to digital objects may be present at any level of hierarchy

The participants noted the need for further work in the a number of areas, including the following:

- formalizing definitions and inclusion guidelines for detailed data elements for the description of digital images and intermediates
- investigating the need for formal cluster-relationship definitions (e.g., naming or characterizing levels of hierarchy, identifying clusters as being for originals, intermediates, or reproductions)

An implicit assumption of the cataloging model developed by the Project was that an amalgam of existing cataloging standards and practices would provide an acceptable formula for the cataloging of digital image collections. The emphasis was on extrapolation from existing cataloging models and incremental modification of current practice. During the final seminar, it was pointed out that it might be worthwhile to develop a more explicit functional analysis for the type of cataloging being proposed, along with a more formal data dictionary, presumably abstracted from AACR2, USMARC and related standards. Whether such a broad analysis is desirable to do at this time, or whether a more careful documentation of known assumptions and requirements will suffice, needs further discussion--particularly in light of parallel work going on in such projects as the Museum Educational Site Licensing Program.

SYSTEM AND NETWORK ACCESS CONSIDERATIONS

National Guide Record. One of the axiomatic assumptions of DIAP was that RLG image collections would continue to require some high-level cataloging or guide record in the national union databases (RLIN and OCLC) in addition to whatever was present at the local level. It was also assumed that local systems might well have more detailed information about collections and items and that links to the actual image files themselves would probably stored and managed at the local level.

USMARC Compatibility of Local Records. Participants agreed, and tested the assumption, that it should be possible to convert from USMARC into whatever format was used locally for the description of image collections. The group did not feel, however, that it could impose "backward compatibility" by requiring that all descriptive and access data stored locally should necessarily be convertible directly back into USMARC. This decision was made, in part, so as not to constrain the exploration of access at the local level to what is currently supported in USMARC. It was also an acknowledgement that local, detailed information may inevitably reside in a variety of different local systems, some of which may simply not be capable of backward convertibility.

This issue naturally raised the important question the of whether it is acceptable bibliographic practice to store definitive descriptive information for collections and items at the local institution that is not only not present in national database systems, but is stored in a format incompatible with existing USMARC-based automated systems. The participants did seem comfortable with this principle so long as navigational tools would be implemented for moving seamlessly between national and local systems, e.g. using URLs stored in national bibliographic records to readily access locally stored data and the images themselves.

This kind of national navigational system would, however, require that local data be stored in some recognized, standard format, if not USMARC. SGML continued to be suggested as the likely possibility, in view of the work proceeding under the aegis of the Berkeley Finding Aids Project. It may be that SGML editing, display, and retrieval tools will become widespread enough in the near future to allow a consistent and standard approach to the creation of detailed local information.

USMARC and Image Access Cataloging. DIAP did not address the possibility of further extending USMARC or AACR2 to accommodate the type of data structure needed for local access to digital image collections. Early discussions in the group brought some degree of consensus that it was probably not useful to constrain local retrieval and access by the current capabilities of USMARC. Work on the inclusion of the URL (Uniform Resource Locator) within USMARC was well under way during this project, however, and was considered an essential key to a national image access system.

The chief drawback to extending USMARC for access to collections of digital images was considered to be its lack of hierarchy and intrinsically flat record structure. Embedding logical hierarchies within USMARC has been attempted in only minor ways ($3, $6, $8) with highly limited functionality and arguable success. In addition, the known problems of USMARC- based systems in handling and displaying extremely long records was considered a major hurdle, particularly in light of the need in some cases to include pointers within a single MARC record to hundreds or even thousands of digital images.

Finally, some project participants recognized the potential problems that multiple versions of digital and non-digital images (e.g., multiple intermediates, images at multiple resolutions, differently edited images, blow-ups, zooms, etc.) would immediately present to MARC, and were reluctant to pursue this issue in the MARC standards arena yet again, in light of earlier unsuccessful attempts to solve the multiple versions problem with respect to the much simpler problems presented by microform reproductions.

During the course of the Project, the RLIN local field 789 (Component Item Field) was in the process of being discussed at MARBI for possible inclusion in USMARC. This field was originally defined for providing linkage from online RLIN records to analog images stored on videodisc as for the Columbia University Libraries Aviador Project. Whether or not the use of this field becomes more widespread, it is clear that it cannot be used adequately to provide more than caption-level access to digital images at a single hierarchical level reflecting a simple parent-child relationship and a single image "generation." This is because it lacks the full complement of descriptive and access-oriented data elements, the capability of reflecting hierarchies, and any technique to identify or distinguish multiple versions or unpredictable component part relationships. Beyond these, the considerations of record length mentioned above come into play. For these reasons and others, it was not considered a

potential solution to the broad set of issues identified by DIAP.

Access System Architectures. A starting point for the DIAP project was the idea of modeling access and retrieval of image information within a local system. The Visual Photologue software, designed by staff at Stokes Imaging, provided an environment that allowed some prototyping work in interface design, indexing and retrieval. This approach had only limited success for a variety of reasons--chief among them the paradigm shift triggered by the sudden more general availability of technology to store images and image metadata locally in open systems and access them nationally and internationally through the Internet and World Wide Web. This new model of distributed access and accompanying software "tool set" was seen as much more promising than the closed, proprietary Visual Photologue system and more in tune with the innovative and well-funded areas of information technology springing up around the Internet.

Nonetheless, the data access architecture implied in both types of systems was that of locally stored detailed metadata and images, linked to (perhaps only summary) metadata in a national union database, and accessible broadly to authorized users via the Internet. Whether the images available on the Internet were to be considered index images and pointers to the locally held originals, or whether the images would be good enough quality to serve as surrogates for some range of uses, was a secondary and evolving consideration.

FIGURE 4: National 'Guide Record' and Locally Stored SRCs

This general access model could of course also be implemented in other ways, for example, with the storage of local information centrally in conjunction with a national union database, or stored redundantly at reflector sites to distribute access load and improve performance.

FIGURE 5: National 'Guide Record,' Distributed SCRs and Images

SGML CATALOG RECORD PROPOSAL

Unlike the Stokes model with which DIAP began, an expanded distributed access model requires that locally stored data, images, and linkage information be represented in a consistent way, following national and international standards to the extent feasible.

Standards for image file formats and for Uniform Resource Locators (URLs) and Uniform Resource Names (URNs) are being adequately addressed in the broader information community at this point. The remaining area, namely, content and format for locally created metadata, is the one most in need of standards work.

Since USMARC is not an adequate tool to encode detailed information at the local level for digital image collections--for the reasons enumerated above--the obvious alternative candidate format is SGML, which is already increasingly being used not only for the encoding of texts but for the representation of metadata. The draft Document Type Definition (DTD) for a USMARC SGML record, prepared at U.C. Berkeley, is a clear starting point for current and future efforts to make USMARC functionally interoperable with the SGML and WWW worlds.

In view of this, the approach proposed here entails the provision of:

- summary records for image (and other) collections in both online national union catalog and local USMARC-based online catalogs
- detailed SGML catalog records (SCRs) for the same collections stored and maintained by the local institution, optionally with thumbnail images incorporated
- multiple image sizes and resolutions stored at the local institution, linked to local SCRs
- pointers (URLs, URNs) from national online union catalog records to detailed SCRs
- local workstation access to summary USMARC data, local SCRs and local images using, for example, a World Wide Web browser that is both Z39.50- and SGML-capable

The proposal also has direct implications for the cataloging of non-collection digital items, particularly reproductions of existing monographic items. For example, a USMARC record for a printed text might well have a pointer to a local SCR for an electronic version of the text, which itself provides access to straight ASCII, TEI-encoded or bitmapped text files along with multiple image files (for illustrations)

including thumbnails and higher-resolution images. This model addresses both the multiple versions problem inherent in electronic reproductions of texts as well as the component part problem presented by multiple text and image files, all corresponding to a single original work.

The approach proposed here has the additional benefit of helping position research libraries to take advantage of the dramatic wave of innovation in access, display, indexing, retrieval and multimedia integration now washing over the Internet. Because library automation is such a small market and relatively poorly-funded, it is neither reasonable nor financially prudent to try to modify existing MARC-based systems to keep up with newer standards and technologies, at least in every case; nor, on the other hand, can we afford to shackle libraries to older MARC- based technology when events are moving so quickly in the world at large. Existing MARC- based systems do their current jobs relatively well and will continue to be crucial as national and local guides to local holdings.

In the developing research library information environment we can anticipate that users will come to access library-related information and documents through a variety of pathways: via RLIN or OCLC, the local OPAC, locally indexed WWW and SGML documents, courseware, electronic publications, remote links and references to local holdings, and many others. MARC- based systems, with their controlled vocabulary and powerful, if rudimentary, search capabilities may provide the most consistently high-quality access to information referenced in their files. They are, however, only one means by which users will find out about digital collections and texts; in many cases users will be led from "courseware," from citations in other electronic documents, or from archival finding aids directly to electronic documents. Integration of local and remote resources will take place at the workstation, through browsers that support WWW, SGML, Z39.50 and other standards. Higher-level "metaindexes" and intelligent "agents" will be needed to navigate disparate information types.

FIGURE 6: Integration of Campus and Remote Information Sources

Note: There was no opportunity to prototype or test the SCR model during the course of DIAP. It is currently being evaluated further at Columbia University Libraries. Operational solutions being studied for the creation and maintenance of SGML Catalog Record and parallel USMARC collection-level records include the initial preparation of an SCR by a cataloger directly into a unix-based system using a DTD-aware SGML editor, followed by the automatic filtering and conversion of that record for loading into local and national MARC-based systems. Upon creation, the SCR would immediately be accessible as a WWW resource by the use of an SGML- capable WWW browser or by local conversion from SGML to HTML on the fly. Additions and changes to the SCR would trigger a job stream that resulted in the corresponding USMARC record(s) being replaced/overlaid with the update record.

APPENDIX A: SGML Catalog Record (SCR) Model

1. The SGML Catalog Record (SCR) would have the following formal characteristics:

- USMARC-based DTD, with high degree of convertibility into and out of USMARC
- capability of incorporating thumbnail images
- incorporation of URL/URN links to external objects
- repeatability of all data elements
- hierarchical data representation
- flexible inclusion of description and linkages to parallel versions and component parts

2. The implementation of the SCR for the cataloging of digital image collections would entail the following:

- Use of USMARC data elements, where applicable
- Use of AACR2 cataloging syntax, where applicable
- Use of AACR2 chapter 13 as a point of departure for the representation of items requiring multi-level description
- The inclusion of all relevant cataloging information, from the summary level to the most detailed level provided; at the summary level, the SCR should virtually duplicate the USMARC record present in the local MARC system and national online union catalog.
- The avoidance of coded USMARC data elements (e.g., the 007), where feasible, and their replacement with normalized, structured-English equivalents
- The use of a simple set of definitions to characterize the logical relationships of parallel and

hierarchical component record parts (e.g., for parent-records, child-records, sibling-records, etc.

APPENDIX B: Schematic illustration of SCR (SGML Catalog Record)

Notes:

1) The SGML tags below are illustrative and do not reflect the draft USMARC DTD. The tags are generally self-explanatory, except perhaps:

- "record" = defines overall record boundaries
- "subrec" = defines subrecord boundaries and includes hierarchy/sequence designator
- "GMD" = AACR2 general material designator
- "URL" = Uniform Resource Locator (or equivalent SGML linking data)

2) The numeric designations following the 'subrec' element in this schematic reflect only hierarchical level and sequence; they have no intrinsic meaning. The need for subrecord relationship designations should be explored.

3) The issue of document structure, e.g., whether a single, long document is created for such a collection or whether a set of hierarchically linked documents is created needs to be explored.

4) The spacing and indentation in the example below are for purposes of illustration only.

EXAMPLE : SGML Catalog Record (SCR) for a slide set; some slides have been digitized.

```
<record>
<titleStmt>
<title>[Columbia University Art-Humanities Image Set] </title> <GMD> [graphic] </GMD>
</titleStmt>
<extent>500 slides : Kodachrome 64 daytime ; 35 mm.</extent>

<subrec>1
<titleStmt>
<title>[Parthenon views]</title> <GMD> [graphic] </GMD>
</titleStmt>
<extent>100 slides : Kodachrome 64 daytime ; 35 mm.</extent>
<subject>Parthenon</subject>
</subrec>

<subrec>1.1
<titleStmt>
<title>[Detail of north pediment]</title> <GMD>[graphic] </GMD>
</titleStmt>
<date>1994</date>
</subrec>

<subrec>1.1.1
<titleStmt>
<title>[Detail of north pediment]</title> <GMD> [computer file] </GMD>
</titleStmt>
<date>1995</date>
<note>[file size, resolution, compression]</note>
<URL>http://www.columbia.edu/cu/arthum/95.32345.gif</URL>
</subrec>

<subrec>1.2
<titleStmt>
<title>[Detail of south pediment]</title> <GMD>[graphic] </GMD>
</titleStmt>
<date>1994</date>
</subrec>

<subrec>1.2.1
<titleStmt>
<title>[Detail of north pediment]</title> <GMD> [computer file] </GMD>
</titleStmt>
<date>1995</date>
<note>[file size, resolution, compression]</note>
<URL>http://www.columbia.edu/cu/arthum/95.32346.gif</URL>
</subrec>

</record>
```