Digitization and Public Access Findings and Recommendations Report

Presented by
Victor Benitez and Valerie Miller
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

28 December 2011
Final Report

Prepared for
The Digitization and Public Access Project
United States Copyright Office
Library of Congress

Table of Contents

STATEMENT OF THE PROBLEM

The United States Copyright Office at the Library of Congress is digitizing the paper records of copyright registrations and concomitant transfers and assignment of rights recorded between 1870-1977. "The records consist of finding aids in the form of catalog cards (~49 million), published volumes of catalog entries (~450,000 pages), source records in the form of applications and staff transcriptions of requests to register works in intellectual property (~26,170 volumes containing ~20 million pages) and copies of documents describing transfers and assignment of rights (~490,050 pages and ~350,000 frames of microfilm)" explains Michael Burke, digital projects planning manager. These records exist in a variety of formats, and a range of sizes and bindings such as preprinted forms on card catalog indexes completed by hand and completed by typewriter, log books, and bound parchment registrations. The variances in appearance reflect the variances in the information recorded in each document as further discussed in the document analysis section of this report.

The Copyright Office is digitizing this one-of-a-kind set of national records not only for preservation purposes but also to make it possible for the copyright records, which are public records, to be accessed from beyond the Washington, DC, office. One of the early and significant project decisions of the Copyright Office digitization and access group was to use digital images of the records rather than model the data in a markup language or attempt to represent all of the data elements in a database. This is an important decision given the scale of the project, the typical conundrums of resource limitations and the need to create a high quality digital collection. What is the best way to make millions of scans searchable? What is the best approach for indexing such a diverse collection? What steps if any, could the project take to produce something more efficient than what has been achieved with the paper collection, or at least to ensure that capabilities are not lost?

SCOPE OF THE INTERNSHIP

The recorded documents have tremendous value in aiding users in the research of intellectual property ownership of registered copyrighted works. This report is the work of two library and information science graduate students selected by the U.S. Copyright Office to work 10-12 hours a week over the course of four months to study the pre-1978 records, analyze indexing options, and consider what elements would be essential to a user interface. It was agreed that in order to make the scope manageable, the study would be limited to the records of recorded documents, with a focus on the documentation of transfers and assignments of copyrights. The interns were tasked with researching the records and recommending appropriate tools and methods for indexing and to prepare a report of findings and recommendations.

CURRENT DIGITIZATION EFFORTS

Overview of *Catalog of Copyright Entries*

The *Internet Archive*[1] began scanning and indexing through optical character recognition (OCR) the *Catalog of Copyright Entries* (CCEs) as a way to preserve and provide further access to the copyright registration records. This is an important step towards responding to user needs as the CCEs are often the first step towards an investigation. OCR of the CCEs, the only complete set of which is held by the Copyright Office, is possible at minimal costs while improving access in a technically simple method. As researchers and Copyright staff know, the CCEs do not include entries for assignments or other recorded documents making them insufficient for a complete investigation. [2] In addition, the *Internet Archives'* does not include creating descriptive interoperable metadata; therefore, a project focused around the recorded documents complements these efforts while setting a blueprint for digitizing and indexing the records as a whole. It is in the Copyright Office's interest to use the *Digitization and Public Access* project to strategically respond to users and contribute toward a semantic web, as several of the titles found in the documents remain under protection of U.S Copyright Law. The project is an opportunity to create good interoperable data that is discoverable on the web and used by a variety of users, such as researchers, universities, digital libraries, and corporate entities.

Need for a digital index of the recorded documents

The *Assignment and Related Documents Index* represents a smaller group or records in a more consistent format and is the only source for tracing documents that clarifies the ownership of copyrights.[3] These unique characteristics make the recorded documents a good candidate for prototyping the digitization and public access goals of the Copyright Office versus dealing with the entire copyright records as a whole. At the moment, the files to the index are in card format and filed through three methods: by title of the work (a total of 1.6 million 4x6 cards in 955 card catalog drawers[4]) by "assignor-transferor" name (83,000 3x6 cards in 40 drawers) and by "assignee-transferee" name (87,000 3x6 cards in 42 drawers). Title cards are found for the period between 1928 through 1977, Name cards were organized separately between 1870 to August 15, 1941, and then interfiled together until 1977 (516,000 4x6 cards in 356 drawers). Cards contain the number assigned to the recordation and the specific volume/microfilm reel and page/frame number. This information points to the corresponding content in the record books in bound volumes 1 through 950 or on microfilm. Tiff images for these

---

[1] The Internet Archive, *Copyright Records,* <http://www.archive.org/details/copyrightrecords>.

[2] United States Copyright Office (Rev 11/2010), *Circular 22: How to Investigate the Copyright Status of a Work*, accessed 12 Dec 2011 <http://www.copyright.gov/circs/circ22.pdf>.

[3] United States Copyright Office (Rev 05/2009), *Circular 23: The Copyright Card Catalog and the Online Files of the Copyright Office*, accessed 12 Dec 2011, http://www.copyright.gov/circs/circ23.pdf.

[4] Digitization and Public Access Project (14 April 2010), *The Copyright Records Digitization Project, Phase 3: Imaging and Indexing*, PPT presentation.

cards (see appendix 3) are already available, imaging was done by DataBank IMX[5], but remain inaccessible to the public.

Overview of digitization workflows and processes

The imaging stage (see appendix 4), currently underway, is the first phase of the digitization and public access project. During this state, cards were selected, prepared, and scanned into uncompressed tiff color image files. The tiff images are stored in a "bag" configuration created under the BagIt specifications as per the California Digital Library (CDL) at the University of California. The CDL sets forth that "BagIt is a hierarchical file packaging format for the exchange of generalized digital content. A "bag" has just enough structure to safely enclose descriptive "tags" and a "payload," but does not require any knowledge of the payload's internal semantics. This BagIt format should be suitable for disk-based or network-based storage and transfer."[6] Naming conventions (see appendix 5) for both tiffs and bags were also created in this stage. Derivative access copies in both JPEG and JPEG2000 color image files in 50:1 compression are being created. The JPEG format is supported by a variety of browsers while the JPEG2000, a lossless compression format, requires a reader like IrfanView to be read by browsers. The overall smaller size of both JPEG formats ensures retrieval and delivery of these images without any strain on library servers.

The second phase of the project is the indexing stage and is currently in the definition and prototyping period. The digitization and access group identified the following minimum descriptive metadata: title, authors, claimants, assignors, assignees, registration/document numbers, date of registration/recordation/execution, links to images of cards/applications/documents. As a prototype, Index terms are being keyboarded by metadata specialists into a Microsoft Access database. As Michael Burke explains, the vision is to capture the data in multiple passes (keyboarding, keyboarding, arbitration) as a way to ensure the highest level of accuracy in data capture. Research is underway to incorporate a double-blind data capture tool with an arbitration pass to ensure quality assurance (see appendix 6).

The digital access management state is the third phase of the project. During this stage, images will be transferred to library storage, derivatives and backups will be created, and unique resource identifiers, registered in a handle server database, will be generated. The fourth and final phase is the integration stage. In this stage, all the components of digitization process will be linked and an interface will be created for public access (see appendix 7 for a full visual representation of all the digitization workflows).

---

[5] *DataBankIMX*, http://www.databankimx.com/.

[6] California Digital Library (CDL), University of California (last ed. 6 Oct 2011), *BagIt File Packaging Format*, accessed 14 Dec 2011 https://wiki.ucop.edu/display/Curation/BagIt.

EVALUATION OF THE PRE-1978 RECORDED DOCUMENTS

Creators of intellectual work often enter into agreements that affect their ownership rights. To better protect their rights, creators may document agreements regarding copyrights with the U.S. Copyright Office as "the law encourages document recordation by conferring certain legal advantages, including priority between conflicting transfers and 'constructive notice'," (Circular 12, 1)[7]. Recorded documents have legal authority before the court of law, as "the Copyright Office maintains a true and accurate copy that can be accepted by a court of law as authentic evidence of the original." The assignment or transfer of rights was recorded for more than 1.7 million works before 1978. The index to these records are only in paper and microfilm format complicating an investigation of the ownership and copyright status of many titles[8] as one of the student's independent search demonstrates below. As the digitization and access group asserts, the loss of this data would exacerbate the problem of orphan works.[9]

Today "Copyright protection subsists from the time the work is created in fixed form," regardless of whether or not the work is registered.[10] This was not always the case. The registration records in this project cover works created prior to 1978. "Before 1978, federal copyright was generally secured by the act of publication with notice of copyright … Federal copyright could also be secured before 1978 by the act of registration in the case of certain unpublished works" (Circular 1).

An example of an independent search to determine copyright status

The graduate students completed guided tours of the copyright records in with Michael Burke and subsequently conducted independent searches. One student worked on determining the copyright status for "Side by Side," a musical composition. The public records of the copyright card catalog in the Copyright Office at the Library of Congress suggest the following series of events.

Copyright for the song "Side by Side" was claimed by Shapiro, Bernstein & Co. from Harry Woods who wrote the music and words. This claim occurred on March 25, 1927. Two copies of the song were received by the Copyright Office. The song was documented as Published. A registration number, R135855, appears in pencil on the title card, which is filed in the "Music" and "Title & Composer" collections from 1898-1937. The class, volume, and page numbers indicated are CL, E, xxc N., 6388E.

---

7 U.S. Copyright Office (Rev 02/2009), *Recordation of Transfers and Other Documents*, accessed 14 Dec 2011, http://www.copyright.gov/circs/circ12.pdf.

8 Michael Burke (8 Dec 2011), "Who owns the copyright for that book, song or photo you want to use? — Making pre-1978 Copyright Office records more accessible" in *Copyright Matters: Digitization and Public Access,* accessed 9 Dec 2011, <http://blogs.loc.gov/copyrightdigitization/2011/12/who-owns-the-copyright-for-that-book-song-or-photo-you-want-to-use-making-pre-1978-copyright-office-records-more-accessible>.

9 U.S. Copyright Office (4 Nov 2011), "Copyright Digitization and Public Access [CDPA]."

10 U.S. Copyright Office (Rev 08/2011), *Copyright Basics*, accessed 12 Dec 2011, http://www.copyright.gov/circs/circ01.pdf.

Prior to 1978, the copyright term was 28 years and renewable during the 28th year. To keep copyright, renewal was required. Therefore, it is probable that copyright was renewed in 1955. On a title card in the "Assignment and Transfers" card catalog collection, "Side by Side" appears again: Date of Execution February 4, 1955, document received February 14, 1955, Class R, No. 135855 of 1954, recorded volume 925, pp 264. It seems copyright for "Side by Side" was indeed renewed in 1955. "Works originally copyrighted after 1922 and renewed before 1978 [have a term that extends] to 95 years from the end of the first year in which they were originally secured" as established in the Copyright Act of 1976.[11]

Apparently a portion of the copyright was reassigned between 1952 and 1958. This is evidenced by a third title card that exists for "Side by Side" in the Assignments and Transfers card catalog, noting that rights were assigned from Harry Woods, Livia Nye Woods and Barbara Woods to Shapiro, Bernstein, & Co., Inc. Details include: Recorded volume 1005, p 361-362, document received April 9, 1958, date of execution April 11, 1952. The fact that there are two assignment cards indicates that different sets of rights may have been transferred. Did the author keep a portion of the rights all along? This may or may not be detailed in the actual assignment and transfer documents, but this exploration ended with the card catalog index cards. By understanding the content, structure and relationships between the sets of copyright records, it is possible to form an understanding of what information is most valuable to future users of the records collection.

Document analysis

In any digitization project, a thorough analysis of the paper records is a logical and critical step toward understanding the documents' uses and determining its most essential characteristics. The interns invested time in understanding the content of the types of records by examining them first hand. The Copyright Records between 1870 and 1977 include registration, assignment, transfer record and associated index cards for literary works, published works, musical compositions, sound recordings, and motion pictures and more.

Over the 107 year time frame there are variations in the information recorded for copyright transactions. This complicates standardization of the terms for the indexing of the surrogate digital records. Although registration numbers were assigned to works granted copyright, these registration numbers were not required for subsequent transactions such as renewals and transfers. Nonetheless these copyright records are acceptable proof of ownership in a court of law. "A transfer of copyright ownership is an assignment, mortgage, exclusive license, or any other conveyance, alienation, or hypothecation of a copyright or of any of the exclusive rights comprised in a copyright, whether or not it is limited in time or place of effect, but not including a nonexclusive license" (Copyright Law, Title 17). The transfer and assignments records consist of legal documents—signed applications detailing the

[11] U.S. Copyright Office (Rev 08/2011), *Duration of Copyright*, accessed 12 Dec 2011, http://www.copyright.gov/circs/circ15a.pdf.

nature of the transaction and the parties and works involved—currently bound and preserved, and associated sets of indexed card catalogs that provide basic information and serve as finding aids.

Through an in-depth document analysis we can begin to predict what researches will want to ascertain from the records, what elements they are likely to search for and by, and thus what indexing terms will serve the project best. The presence of seals, types of signatures and dates, the listing of parties, the transaction descriptions, in this case, point to the fact that these copyright records are legal documents addressing ownership.

The digitization project descriptions indicate that there are 36 types of source records. Table 1 (see appendix 1) shows a majority of the fields that can be culled from an assignment and transfer document. Limitations on time and resources make it impossible to use a markup language to model each paper record in full. The Copyright Office's OCR attempts produced a high level of inaccuracy. Elements that are useful in a deep analysis or even a typical records search, but which are the least likely to function as search parameters, have been listed in Table 2 (see appendix 2). These items are best read directly from the copy of the document record. Details such as the fact that the index card is from a printed series of stock cards created after a change of copyright was made effective according to the dates on the same index card can be valuable in an analysis of the records, but are unlikely to be useful in an actual search. This is the type of information that falls into Table 2. The document analysis assists in prioritizing available data.

USABILITY FRAMEWORK

Once the designated documents have been examined, it is appropriate to use the information to formulate conceptions of how future viewers will use the digitized collection. Scholars and field experts have explained that usability concerns should be explored and incorporated into a project from the design stage. If an interface is not user friendly, if users have to wait longer than usual or find it inefficient in meeting their needs, they simply click elsewhere explains usability expert, Jakob Nielsen, Ph.D.[12] Adopting techniques like wireframing, mockups, and Personas and Scenarios helps conceptualize objectives and ensure resources are being most efficiently directed towards producing an interface that meets expectations of those the project wishes to serve.[13]

The Personas and Scenarios model is a user-centered approach that can make a positive impact and produce measurable success. It will be described here because an organization can implement this technique with knowledge and a consistent commitment to be effective. It requires no additional funding. This technique requires the creation of three to six archetypal users. The imaginary characters

[12] Nielson, J., *Usability 101: Introduction to Usability*, accessed 5 Oct 2011, http://www.useit.com/alertbox/20030825.html.

[13] Kutz, M., Miller, V., Suber, S. (2011), *Usability and Design Info Pack*, https://sites.google.com/site/udinfopack/personas-and-scenarios.

are named, assigned hobbies, goals, and a set of needs and preferences related to use of the digital interface. These Personas and Scenarios help the stakeholders and interface builders synchronize project priorities.[14] The technique provides guidelines by which decisions can be made.

For example, the digitization project might conceive of a user named John Salisbury who works for a publishing company researching copyright status of titles it has indirectly acquired. Salisbury is an experienced user of the copyright records, he prefers to explore as many angles as possible before making definitive conclusions. He expects to make more than one search to avoid ambiguity, and is patient about sorting through results. Although precise search results are good for Salisbury, he is most confident when reaching conclusions as a result of information unearthed from his own search results. Therefore, if a collection is built and indexed around the needs of Salisbury it would be different than one created based on the needs of other types of researchers considered. More specifically, searching functions that produce more results versus more precise results may be prioritized for Salisbury because he is perfectly willing to comb through results and drill down. Salisbury would be frustrated with a search that produces narrow results and does not provide leads to collocated records and allow him to browse. That is an example of the type of frameworks for decision making that emerges when the Personas and Scenarios method is effectively utilized.

METADATA OVERVIEW

Library Science experts explain that Administrative Metadata includes information about ownership, custodianship, access rights and management. Collectively, the Copyright records constitute the Administrative Metadata for a work. A comprehensive collection of this metadata is not held by any organization other than the U.S. Copyright Office within the Library of Congress. The information used for sorting, searching and accessing these records, such as date, authorship, and title, fit into the category of Descriptive Metadata. Any metadata used in the ongoing preservation of the digital records can be categorized as Preservation Metadata (see appendix 8 for diagram version 2).

Structural metadata

The Metadata Encoding and Transmission Standard (METS) is a schema for encoding descriptive, administrative, and structural metadata regarding objects in a digital library and is expressed using the XML schema langauge.[15] The descriptive section is followed by an administrative section where technical metadata is placed, usually in the PREMIS standard,[16] for the digital surrogate image. Administrative metadata created during the creation of the digital surrogates could be used and dropped into this section. The structural section is represented by the file section, structural map, and

[14] Adlin, Tamara (2010), *The Power of Ad Hoc Personas: Truly Practical Methods to Get Your Organization on the Same Page* (PowerPoint), *UIE.com*., accessed 16 Oct 2011, http://www.uie.com/events/virtual_seminars/ad_hoc_personas/.

[15] *Metadata Encoding and Transmission Standard*, http://www.loc.gov/standards/mets/.

[16] *Preservation Metadata Maintenance Activity,* http://www.loc.gov/standards/premis/.

structural link sections and ensures all objects represented in the metadata are mapped correctly without jeopardizing the integrity of the objects. The structural section uses XQuery language through XLink attributes points to the digital surrogates upon retrieval. If adopting METS is impractical during the implementation stage, the structural sections should be analyzed closely as they provided a framework for the implementation of the XQuery and XLink XML languages that help ensure the structural integrity of the digital objects. This section is extremely relevant for creating appropriate pointers in the customized schema will be discussed ahead.

Metadata and interoperability

With metadata, if it is available, other organizations and interfaces will be able to interpret and make use of the collection. This is the trend in collections management. Optimally, there is a seamless interaction between web interfaces. For example, with one click a user can go from a cataloged item at the HathiTrust to a WorldCat list of libraries holding print versions of the item.

Another example of an organization anticipating possibilities for institutional collaborations, is the Brooklyn Museum digital collection which is accessible via API.[17] One intern extracted a collection of items related to a search term and experimented with displaying it in Greenstone a server based library software. An extraction consists of metadata in XML or HTML and may include a stable URI. Numerous established organizations and digital humanities projects feature API's which allow access to their collection, from the New York Times[18] to WeFeelFine.org[19]. NPR.org also offers an API, and describes it as follows:

> An API, or Application Programming Interface, is a way for two computer applications to talk to each other in a common language that they both understand. NPR's API is a content API, which essentially provides a structured way for other computer applications to get NPR stories in a predictable, flexible and powerful way. The content that is available includes audio from most NPR programs dating back to 1995 as well as text, images and other web-only content from NPR and NPR member stations. This archive consists of over 250,000 stories that are grouped into more than 5,000 different aggregations.[20]

Contemporary expectations of digital collections, particularly those held in the public interest, are that they be shareable, interoperable and allow for collaboration. Efforts at standardization of metadata are massive and international, engaging numerous organizations from the W3C, Dublin Core Metadata Initiative, and JISC to the Library of Congress, just to name a few. Whether RDF, XML or a standard relational database is in use, metadata keeps your information from becoming isolated, it in fact makes it findable, and usable, which in turn adds value to the collection, so it is critical that metadata be included and that standards be employed whenever possible.

---

[17] Brooklyn Museum (2011), *Brooklyn Museum Collection: API*, accessed 21 Oct 2011, http://www.brooklynmuseum.org/opencollection/api/.
[18] http://developer.nytimes.com/gallery
[19] http://www.wefeelfine.org/api.html
[20] NPR.org (2011), *NPR Tech Center*, accessed 26 Dec 2011, http://www.npr.org/api/index.

Unique resource identifiers

As Michael Burke argues, naming conventions should reflect and easily identify the content of the images. Unique resource identifiers (URIs) used in the final digital record must also easily identify and point the users to the digital image surrogate. A file naming convention has been employed that cleverly captures at least one date, information about the original filed location, the original document type and more. Selected information related to each record is being captured in database fields with plans to use that data for an index. A standard, "rfc 3650,[21]" exists and provides a framework for creating unique identifiers under the handle system. The standard provides insight into the secured name resolution and administration over the web for the handle system and further context of the framework in relationship to other URI standards.

INDEXING OPTIONS

Level of indexing

Debate about the level of indexing is ongoing among the Copyright Office staff. This debate centers on the desire for full-text searching of the copyright documents, which has been expressed by users in surveys and meetings with the public. As library students the desire for full-text searching is understood, but the reality of resources available to the Copyright Office and the technical complexity of achieving this environment outweigh the benefits of this feature. Instead, the digitization and access group should proceed by indexing the catalog cards and incorporating pointers to digital surrogates of the corresponding document copies. Correct mapping between the digital surrogate records and digital surrogate images should be a priority during the implementation stage, which can be done by incorporating a structural metadata standard like METS for which an overview has already been provided earlier.

Privacy concerns

The Copyright Office clearly states that copyright records are public; yet, concerns over privacy issues remain. Recorded documents are "legal documents [that] include wills and contracts transferring copyright from one person or firm to another" (CDPA 11/4/11, 4) and sometimes contain identifiable information. To protect copyright owners' privacy an argument can be made against the full-text indexing of the documents. If full-text indexing were possible, indexing this data would make it discoverable and interoperable in the web. Although this creates transparent and open information, which the interns as library students strongly advocate, trust between intellectual property owners and

[21] Network Working Group (Nov 2003), *Handle System Overview: RFC 3650*, accessed 15 Dec 2011, http://www.ietf.org/rfc/rfc3650.txt.

the Copyright Office may be ruptured. If this trust were ruptured it could adversely affect the Copyright Office's efforts; even though, registration and recordation of assignments and transfers still largely outweigh these issues. In an environment where only the copyright index cards are indexed and digital surrogate images are provided, researchers would need to first locate the records, retrieve their digital surrogate images, and then read the static content of the documents. The fact that the content of the digital surrogate images remains static helps create a barrier of protection for copyright owners' identifiable information.

Considering concerns of privacy and indexing goals, could a Work Title-centric digital object, a digital data record, be created from the copyright registration records and all associated finding aids related to a registered work? The digital data record would be a composite data record consisting of information culled from an assortment of registration records. The objective is for these digitized records to offer new functionalities for searching and new opportunities for interoperability within this system and in relation to others in the Library of Congress as well as at large. The success of this depends on the relationships that are applied to the data elements either as they are stored or as they are extracted. It can be argued that this is the type of functionality that will be expected from the digital records and anything less will have to be carefully explained at the public interface point.

However, as mentioned, time does not allow for staff to make these *narrative* connections about the history of copyright in a work. Professional researchers have estimated that it takes approximately five years of training and experience to become an expert copyright researcher.[22] As of November 2011, the price for a member of the public to engage the expertise of the research staff at the U.S. Office of Copyright was $300. For that cost, a report matching provided search parameters is produced with no guarantees of definitive results. The records are open to the public, but any narrative created from the records is kept confidential and not even retained by the Copyright Office.

XML and Indexing

Using Extensible Markup Language (XML) is an optimum way to allow for facets to be applied to the data. XML is a metadata markup method that allows for customized, machine readable detailing of works presented or represented in digital form. There are numerous accepted, formalized schemas and many to be invented. However, there is a balance to be found in terms of utility and deciding how unique or how interoperable an encoding schema should be. There are formalized guidelines in place. One of the governing organizations involved with XML standards is TEI, the Text Encoding Initiative.[23]

XML is endorsed by the W3C for use in the semantic web. XML is used widely and known for its extensibility and contributions to interoperability. For managing data, the comparable method is utilization of databases. One of the weakness of XML as compared to databases, when handling large quantities of data is that the hierarchical structure impedes quick searching. Indexing the data for

[22] Kelly, Rosemary. Copyright Office Head Researcher. Interviewed in October 2011.
[23] http://www.tei-c.org/index.xml

quicker access is necessary to "achieve in quantity what the Internet demands." That is a phrase used by TEI contributor Julia Flanders[24] to describe the need for automated processes when you have what MarkLogic executives call "Big Data." The 49 million records being digitized by the Copyright Office qualify as big data. Indexing means applying an ordered representation for quickly locating nested information. In the case of XML indexing, the ordered representation is overlaid and linked to the data it describes.

MarkLogic is one of the software solutions that provides indexing and server services. Other comparable services are eXist, DBXML, Xindice, XTF, and Apache Solr.

> MarkLogic Server fuses together database internals, search-style indexing, and application server behaviors into a unified system. It uses XML documents as its data model, and stores the documents within a transactional repository. It indexes the words and values from each of the loaded documents, as well as the document structure. And, because of its unique Universal Index, MarkLogic doesn't require advance knowledge of the document structure (its "schema") nor complete adherence to a particular schema. Through its application server capabilities, it's programmable and extensible.[25]

It is relevant to note, first of all that more than one XML schema can be used to guide the data capture, and design data output on a project. Secondly, with sophisticated tools like MarkLogic, schemas can be applied as needed, not dogmatically.

Framework for creating a customized schema

Implementing a metadata standard for the administration and description of the digital surrogate records is an important step towards creating interoperable data and supporting the semantic web. Current data markup technologies include RDF, XML, Microformats. Each of these has its own set of standards, rules, strengths and weaknesses. RDF for example is endorsed by the W3C and a hot topic for research in the Library Science community. Competing with RDF is the HTML 5 based microformat language.

Both RDF and microformats are used to create linked data by providing semantic relationships in machine readable format, so that inferences can be made, such as this "string of text which is an author's name supplemented by a URI" wrote this "string of text which is a URI for a book." Beyond simply creating semantic relationships between html pages, linked data methodology seeks to create semantic relationships at an elemental level using triples, and simultaneously at a broader level by mapping and defining namespaces and data sets via the URI's connected by the relationship statements.

[24] Flanders, Julia. From a lecture in September 2011.

[25] Popescu, A. (2010), *MarkLogic Server: Data Model, Indexing System, Operational Behaviors*, accessed 5 Oct 2011, http://nosql.mypopescu.com/post/1429730551/marklogic-server-data-model-indexing-system.

By using a URI that refers to a description of a thing and then connecting an object to that initial thing or subject, a triple is created. Any number of objects can be connected to any number of subjects by a predicate. A predicate names the relationship between the subject and object. Two of these elements have URI's and the URI is created from a relevant name space. The third element, the predicate may be derived from an appropriate schema.[26]
Conversion technologies exist for taking records from databases to XML, to RDF and back again. The question is what serves the project needs most efficiently and expeditiously.

Currently, no established descriptive or administrative metadata standard has been found appropriate by the digitization and access group, as mentioned in the section on metadata, these records offer a unique data set. As a result, the creation of a customized schema is needed. Established standards remain relevant as they can be used as a framework and provide insight for schema development. When developing a customized schema important questions need to be addressed. These questions include: who has the authority to create the schema?; who should be involved in the creation of the schema?; how will user needs be reflected by the schema?; and how will the schema be tested?

Relying on the knowledge built in analyzing the documents involved in the digitization, two customized schema (see appendix 9) were created to investigate the feasibility of creating a customized schema. The schema ere encoded in XML format. One student used the RNG rules for markup validation. The second student created a corresponding schema using only XML markup principles and modeled it after post 1978 recordation records found in the Voyager catalog and pointers to images were incorporated based on those found in the PREMIS metadata standard.

An XML Pseudoschema

Appendix 8, diagram version 2 is an illustration of how metadata can be categorized for this project. It also offers an illustration of the RNG based XML schema, diagram version 1 (Appendix 8). The diagram and the XML itself, element by element can be explained, step-by-step:

The root element is Collection and it can contain any number of images. Each Image is an element with two mandatory descriptors, filename and location, associated with it. No record exists without an Image. Every image must have a filename and location. The location is meant to be the URI, a permanent server location used for the image and its metadata, or a permanent link/pointer to that location. Each image contains, in a hierarchical sense, five elements –Work Title, Registration Number, Date, Name, and Action. Each Name contains one element, Relationship. As it is with Image, Work Title,

[26] Davis, I., & Heath T. (2009). 30 Minute Guide to RDF and Linked Data|Slide show on *SlideShare.Net*. Retrieved from http://www.slideshare.net/iandavis/30-minute-guide-to-rdf-and-linked-data
Berners-Lee, T. Tim. (2009). Berners-Lee on the next Web | Video on TED.com Retrieved December 3, 2011, from http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html

Date, Action and Relationship are each associated with their own set of attributes meant to describe the data that fits into the element.

Now that we named each of the elements and their root container, it is easier to further discuss which are mandatory, what the attributes are and when they have to be applied. These rules are listed in the RNG document along with the element definitions. When an XML document is created with an internal reference to the RNG document, that XML document can be determined valid or invalid compared to its RNG rules, by software that reads or writes XML. In this specific example, although it is useful to record every element, in many cases this information is not available. Variations in available information have been anticipated in the design primarily because this schema is meant to be functional for any of the record types associated with the digitization project.

An Image record can be populated in XML with the other five elements in any order and any amount. According to the RNG rules created in this attempt, Date and Action are the only required elements for the Image, meaning every image has to have at least one Date and one Action while it may have zero Registration Numbers or four. This is defined by using tags such as zeroOrMore and OneorMore in the RNG validation document. Reasons for these choices are explained in the Document Analysis section. The Date is set to be acceptable only in ISO format, by use of the data type tag. Using W3C compliant data types makes the records that much more machine readable. The Date can be labeled as an execution date, publication date, date of recordation, date of witness according to signature, the date the document was received, or unidentified. The date must have one type. There can be any number of dates.

For every Name associated with the record a Relationship must be applied. It is worth noting that within the element definition, the RNG document dictates whether within the Name, the first name is mandatory, whether it appears after the last name, and whether or not a middle name or initial is required. Keeping in suit with sample copyright records these fields are set to be fairly flexible to allow for instances where only initials are provided for first name and middle, and other such variances. Every Name must have a Relationship of AssignorFrom, AssigneeTo, Contributor, Other, or Undefined. These parameters are designed to accommodate the multitude of recordation types and the terminology historically used in the paper records.

Every Action should be labeled Renewal, Transfer or Registration, and these categories are evident in the record type at the point of data entry. More details can be assessed by looking at the RNG document, the sample XML record and the two index cards used to complete the XML record (see Appendix 10). This step-by-step is best understood while looking at the associated schemas and diagrams. Any discrepancies between diagrams and text can be attributed to the evolution of thought as the possibilities were explored during the length of the internship; no concepts are lost within the slight variations.

Conceptualizing in XML

It is useful to approach the project from the view of a systems analyst and use XML to help understand the information relationships between the records and the data they contain, to answer questions like: What would an XML schema for the Copyright records look like? How can current metadata standards be incorporated into the schema for integration with existing data stores and online catalogs held by the Library of Congress, conglomerates like the Hathi Trust, non-profit endeavors like Open Library, and even smaller digital humanities or scholarly research projects.

Assuming privacy concerns could be surmounted, if the records could be connected to allow a search by a work's title to yield the history of copyright of that work, this would mean that the digitized records have not only served to provide access beyond the actual office where the records exist in Washington, DC, but have also improved upon the paper records in a significant way. Imagine the differences between a drawer by virtual drawer search versus browsing and editing suggested histories based on the information the computer interface provides.

When we diagram theworkTitle centric relationship, a record centered around the title of a work, and the actual data as it is currently being captured, it is clear that the XML for the existing scenario and the ideal one are highly similar. See (appendix 8, diagram version 1) the RNG based schema to assist in visualizing this concept. It is conceivable that their similarity indicates that, employing a simple but effective XML schema for recording the data and using it to create a digital data record may be achievable with the same time frame and resources as would otherwise be spent.

As novice practitioners in XML, we can claim that modeling data in XML offers unique flexibility to a digitization project. XML could be used to record information about the original resources in an interpretive way, attempting to capture descriptive features such as whether the originals contained typed or handwritten information, misspellings or official stamps, etc. Given the quantity of records that need to be digitized for the copyright records digitization, this type of detailed modeling is not feasible. XML best serves this digitization project if it is used as a means of recording data for creating not only a faceted index, but also data that can be collocated by elements and attributes, ultimately allowing for more robust results from searches than initially considered possible with the resource constraints. Our work in XML for this project serves to illustrate what it means to create an XML document, and to suggest what needs to be considered.

XML can function as an alternative to databases. Comparing databases to XML, field expert Ronald Bourret states that "a product like a native XML database or a content management system…will allow you to preserve physical document structure, support document-level transactions, and execute queries in an XML query language." [27] Its limitations in terms of being restricted to sequential, hierarchical searching can be nullified with the right management tool, such as MarkLogic or Apache Solr.

[27] Bourret, R. (2005), *XML and Databases*, accessed 11 Sept 2011, http://www.rpbourret.com/xml/XMLAndDatabases.htm.

With effective implementation, XML, theoretically, is a technology that could allow users to search the records and construct possible histories of works of interest to them. A user could pull up records that seem to relate to her search and hold them in a "shopping cart." In which case, buying the items would be equivalent to accepting the collected records as a probable copyright history of the work under research.

NATIONAL INFORMATION EXCHANGE MODEL

The National Information Exchange Model (NIEM) is an alternative method for creating XML schemas. In the wave of open data and government transparency, the adoption of NIEM standards ensures data interoperability between agencies at all levels of government, the private sector and international partners.

> [NIEM] presents an approach to driving standardized connections among and between governmental entities as well as with private sector and international partners which enable disparate systems to share, exchange, accept, and translate information. With the use of NIEM framework comes greater agility and efficiency in satisfying business needs and implementing repeatable processes. The common data connections developed using NIEM results in reusable artifacts that reduce future development costs resulting in cost avoidance. (Federal CIO Council, 1)[28]

Through the implementation of NIEM standards, XML schema development is a cooperative process that leverages established knowledge and resources. This process ensures data is "discoverable and standardiz[ed] as it moves in between the current siloed stores across the Government" and the web. As the Federal CIO Council states below:

> Using NIEM as part of a broader data strategy supporting Enterprise Architecture means that the organization has agreed to challenge the "Status Quo" and has started on a path for innovation, light technology, and shared solutions as outlined in the Office of Management and Budget (OMB) 25 Point Implementation Plan to Reform Federal Information Technology Management. NIEM as a tool empowers agencies to create and maintain meaningful data connections across their stove-piped information technology systems as well as across their stakeholder base of other … partners. (2)

Implementation of NIEM in the creation of an XML schema is more than just the acceptance of a particular set of standards, but a broader strategic strategy to ensure interoperable data while keeping costs of development to a minimum as a result of shared resources.

[28] Federal CIO Council (2010), *Agency Information Exchange Functional Standards Evaluation: Adoption and use of the National Information Exchange Model (NIEM), accessed 26 Dec 2011, https://www.niem.gov/documentsdb/Documents/Other/AssessmentReport.pdf.*

NIEM's data model consists of terms agreed upon through a rich governance process of practitioners at all levels of government and private industry. The model is comprised of the NIEM core, NIEM domains and NIEM tools. The NIEM core consists of data elements agreed upon regarding the semantic and syntactic representation commonly understood across all domains. NIEM domains, extensions of the NIEM core, include "mission specific data that is managed through independent stewards" (Federal CIO Council, A-1). As of the publication of the *Adoption and use of NIEM* Report, "ten domains exist ranging from international trade to family services" (14). Future domains are added as necessary based on an established need (A-1). NIEM tools, "a reference set of tools freely available with each NIEM release" (A-8), are used to develop an Information Exchange Package Document (IEPD). An IEPD documents "a discreet information exchange for reuse across a larger community or mission space" (A-1). "The IEPD Life Cycle [comprised of six phases] is a best practice, which defines the steps required to identify and document information exchange use cases and requirements, develop an IEPD, and make it available for search and discovery" (A-2).

A search in the IEPD clearinghouse[29] for schemas that may be relevant to the Copyright data locates the "Certificate of Real Estate Value"[30] IEPD, developed by the Minnesota Department of Revenue, Department of Property Taxation. The IEPD is used for "the transfer of real property between agents with real property fair market value analysis." The schema includes the following elements: docmentid, submitter, buyer, sellers, realproperty, salesagreement, fairmarketvalue, countydata, crvworkflowstatus and contains subelements for personal contact information. Although the schema goes beyond the Copyright's data needs, its similarity may serve as a framework for modeling the Copyright's data.

ANALYSIS OF OPEN SOURCE APPLICATIONS

Apache Solr[31]

Solr, a strong alternative to MarkLogic, is an open source enterprise search platform developed by the Apache Lucene project and implemented by majors players like the White House to Groupon. The interface offers next generation searching capabilities that are highly desirable by users. Among these capabilities major features include advanced full-text search capabilities, hit highlighting, faceted search, dynamic clustering, database integration, rich document handling and is optimized for high volume web traffic. Solr is written in Java and runs as a standalone full-text search server within a servlet container. Solr uses the Lucene Java search library at its core for full-text indexing and search, and has REST-like HTTP/XML and JSON APIs that make it highly customizable using almost any

[29] IEPD Clearinghouse, http://it.ojp.gov/framesets/iepd-clearinghouse-noClose.htm.

[30] Minnesota Department of Revenue Department of Property Taxation, *Certificate of Real Estate Value*, accessed 15 Dec 2011, *http://niem.gtri.gatech.edu/niemtools/iepdt/display/container.iepd?ref=AxkgJWIedJU%3D*.

[31] Apache Software Foundation (2007), *Introduction to the Solr Enterprise Search Server*, accessed 14 2011, http://lucene.apache.org/solr/features.pdf.

programming language. This robustness is what makes Solr particularly useful as it can be incorporated with almost any open source tool like Drupal, Omeka (both content management systems), or Koha (a open source free version of Voyager) and linked to one or multiple open source databases.

The Lucene Search Library is a real data schema with numeric types, dynamic fields, and unique keys. The library comes with a built in administration interface and can be externally configured via XML. The library accepts XML schemas and indexes data without any downtime. Data can also be imported from a databse via a data import handler or loaded from a CSV file. Aside from this robustness, Solr is highly supported by developers who contribute to its extensive collection of documentation in its wiki at the Apache Solr web site.[32] A Google search for Solr reveals everything from a step-by-step implementation how-to video to examples of countless integrations with various open source applications.

Weighing the risks and benefits of open source solutions

Open source solutions must be carefully weighed through a variety of factors. Although the costs of the software are free, hidden costs are seen for labor and time associated with its implementation. A comfort with programming languages is necessary for a successful implementation. Security concerns must be properly addressed or data loss is risked. On the benefits side, open source offers the ability to be highly customizable to the needs of users. An evaluation of support in the form documentation available and directly from developers on the wikis and in the web should be assessed. An assessment of the skills found in the organization looking to implement the software should also be made. This assessment should consider the programing ability of the staff, the development environment, support from information technology and from other departments in the organization.

CONCLUSION

Understanding the U.S. Copyright records and the challenges of digitization is a complex task. This report hopes to bring new points of consideration to the endeavor. One of the most strident insights gleaned in researching the topic is the benefit of making these records not only searchable but also useable for other organizations. Considerations over issues of privacy as a result of the exploitation of public information are required. These privacy considerations need to be balanced against the benefits of providing online public access. Further considerations are also needed over how third-party entities will/may use the resulting data. In the meantime, the fields of library science and information technology at their intersection with data curation point to the value of information as assessable in terms of whether or not the data can be verified as authentic and used to produce new contexts and conclusions. Whatever technologies can be effectively employed to facilitate this interoperability will add to the worth of an information collection.

[32] Apache Solr (last ed. 7 Dec 2011), *Solr Wiki*, accessed 16 Dec 2011, <http://wiki.apache.org/solr/FrontPage>.

The practical considerations of human resources, time and the immense number of records leads to very practical decisions about what data is carried over to the digital surrogate record and how the digital records are stored, named and retrieved. Adding functionality to the records loses priority due to the difficulties in reproducing paper system functionality in a digital system. In the current digital environment, immediate data delivery through effective and efficient means is the expected norm. Data no longer lives in a vacuum, siloed away in a particular information interface constrained to one specific static format, but is dynamic as content and technology are now one. Data encoded in XML makes this possible. In this changing environment, libraries along with non-profit organizations and public agencies must compete against propriety efforts to remain relevant. This is the type of real world conundrum that is most easily addressed through collaboration and participation in information standards or open source communities.

Given the abundance of information standards and a growing community of professional and amateur developers, data modeling and curation should not be done in a vacuum. This is particularly true in the development of XML schemas. Although the Copyright's data is relatively simple as it is based on the paper index records, important questions about schema development remain. This is where the adoption of NIEM is particularly relevant as it forges cooperation, governance considerations and a variety of resources while keeping costs of development to a minimum. As NIEM creates data encoded in XML, it is entirely compatible with Solr as the Department of Homeland Security Digital Library[33] demonstrates. As of June 2010, twelve federal agencies were committed to using NIEM and eight others were committed to further evaluation (Federal CIO Council, 8). Data created under NIEM is not only interoperable but also transparent and open, which has resulted in an abundance of third-party uses and the creation of a variety of applications. As copyright data is extremely valuable in the rapidly expanding digital environment, third-party developers are likely to find uses for the data further improving access alongside the Copyright Office's efforts.

[33] http://www.hsdl.org/

APPENDICES

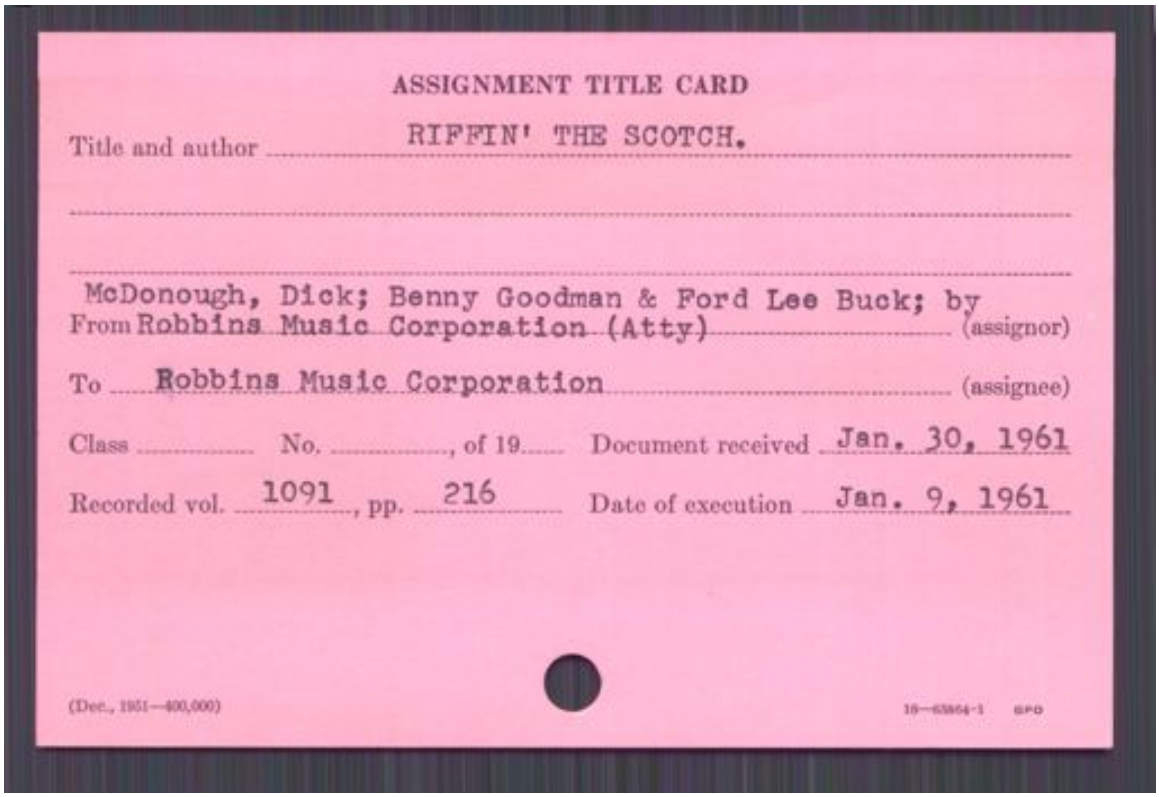
Appendix 1

<table>
<tr><td>Element</td><td colspan="5">Qualifier, Attribute, or Sub-index Element</td></tr>
<tr><td>Title*</td><td colspan="2">Class (indicated by a capital letter representing the category of the work)</td><td colspan="3">Published or Unpublished</td></tr>
<tr><td colspan="2">Assignor / From</td><td colspan="4" rowspan="3">Party/Entity</td></tr>
<tr><td colspan="2">Assignee / To</td></tr>
<tr><td colspan="2">Author/ Creator</td></tr>
<tr><td>Date</td><td colspan="5">Execution, Publication, Recordation, Witnessed, Document Received</td></tr>
<tr><td colspan="2">Registration Number</td><td colspan="3" rowspan="2">Whenever these elements don’t exist for the title, the digital data record should indicate this.</td><td></td></tr>
<tr><td colspan="2">Assignment and Transfer</td><td>Date</td><td>To</td><td>From</td></tr>
<tr><td colspan="6">*Titles are the most consistent element across the resource records. However, no title itself is under copyright. Therefore titles repeat and exist in slight variation within and across classes.</td></tr>
</table>

Table 1

Image Only (Tiff) Elements and features that will not be machine readable
Signatures of parties and officiators, stamps and seals, irregularly appearing data points such as the initials of the indexer or recorder, identifiers from the series of paper cards – numbers, colors, format, number of copies of the work received, amount paid, variances in format of the registration/application books, volume and page numbers of recordation.

Appendix 3—Assignment Title Cards Tiffs



ASSIGNMENT TITLE CARD

Title and author RIFFIN' THE SCOTCH.

McDonough, Dick; Benny Goodman & Ford Lee Buck; by

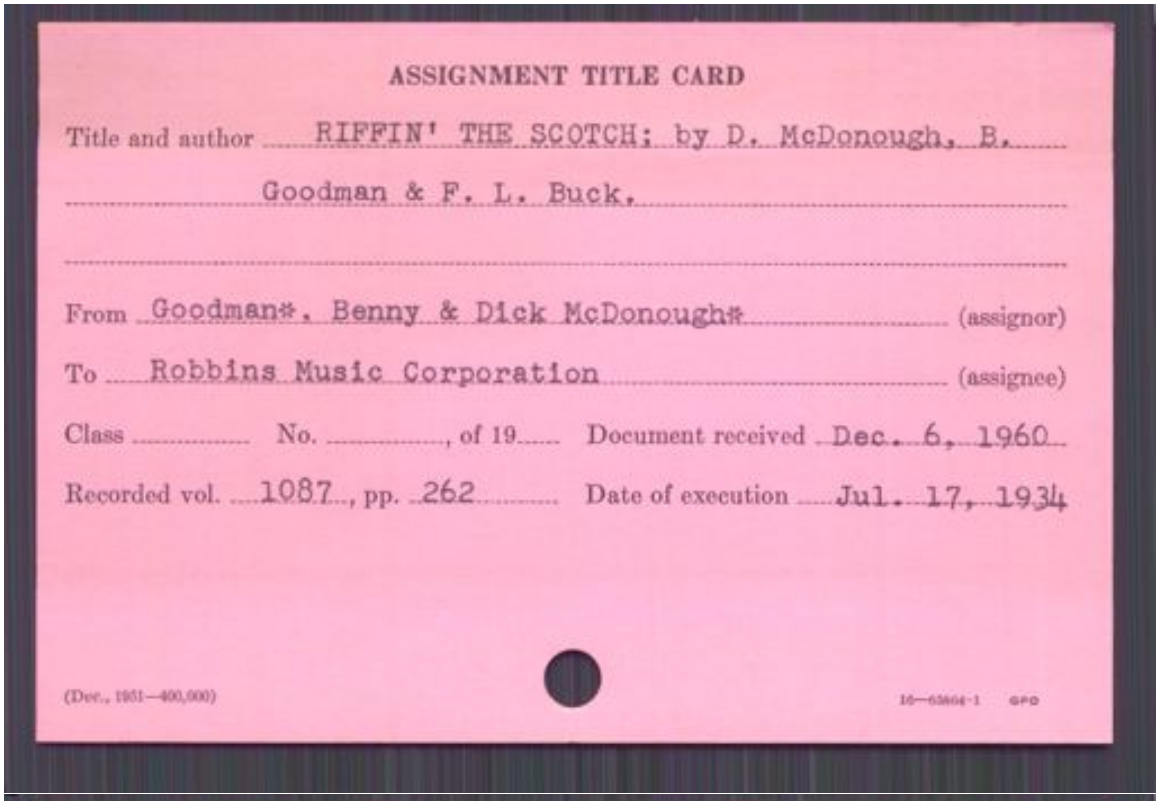
From Robbins Music Corporation (Atty) (assignor)

To Robbins Music Corporation (assignee)

Class ..... No. ....., of 19..... Document received Jan. 30, 1961

Recorded vol. 1091, pp. 216 Date of execution Jan. 9, 1961

(Dec., 1951—400,000) 16—63864-1 GPO



ASSIGNMENT TITLE CARD

Title and author RIFFIN' THE SCOTCH; by D. McDonough, B. Goodman & F. L. Buck.

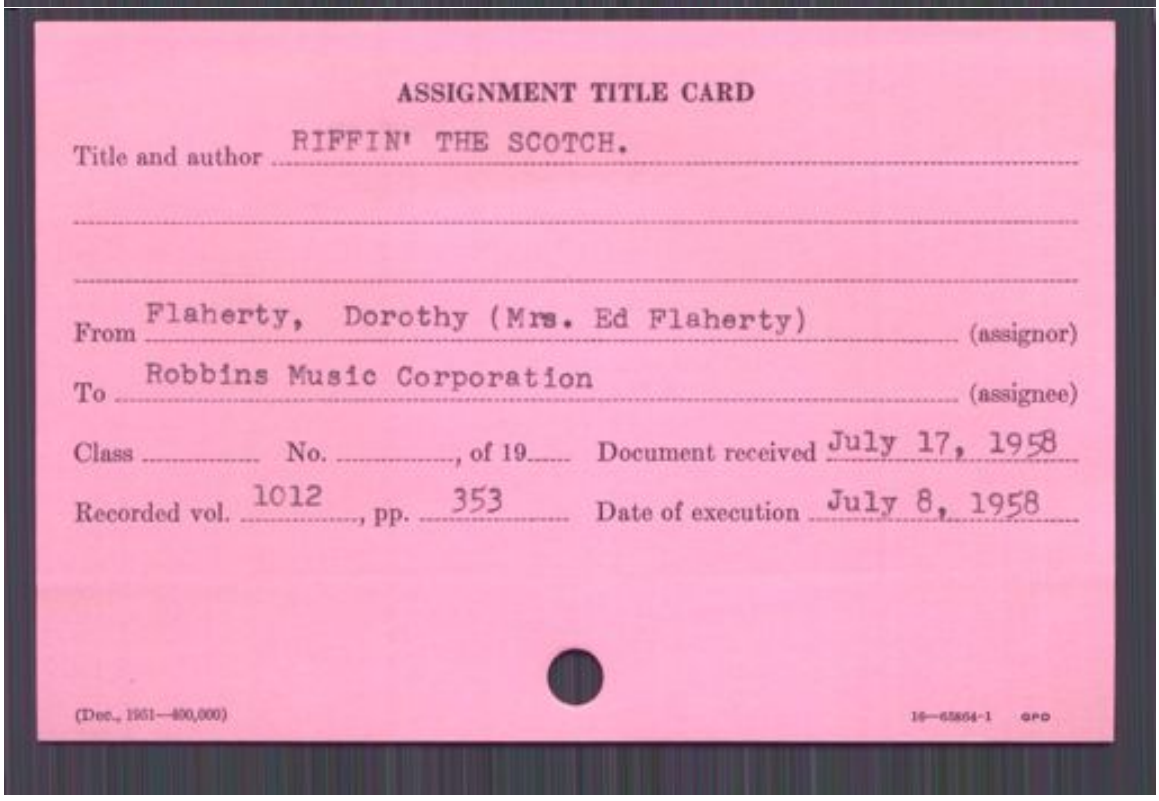
From Goodman*, Benny & Dick McDonough* (assignor)

To Robbins Music Corporation (assignee)

Class ..... No. ....., of 19..... Document received Dec. 6, 1960

Recorded vol. 1087, pp. 262 Date of execution Jul. 17, 1934

(Dec., 1951—400,000) 16—65804-1 GPO



ASSIGNMENT TITLE CARD

Title and author RIFFIN' THE SCOTCH.

From Flaherty, Dorothy (Mrs. Ed Flaherty) (assignor)

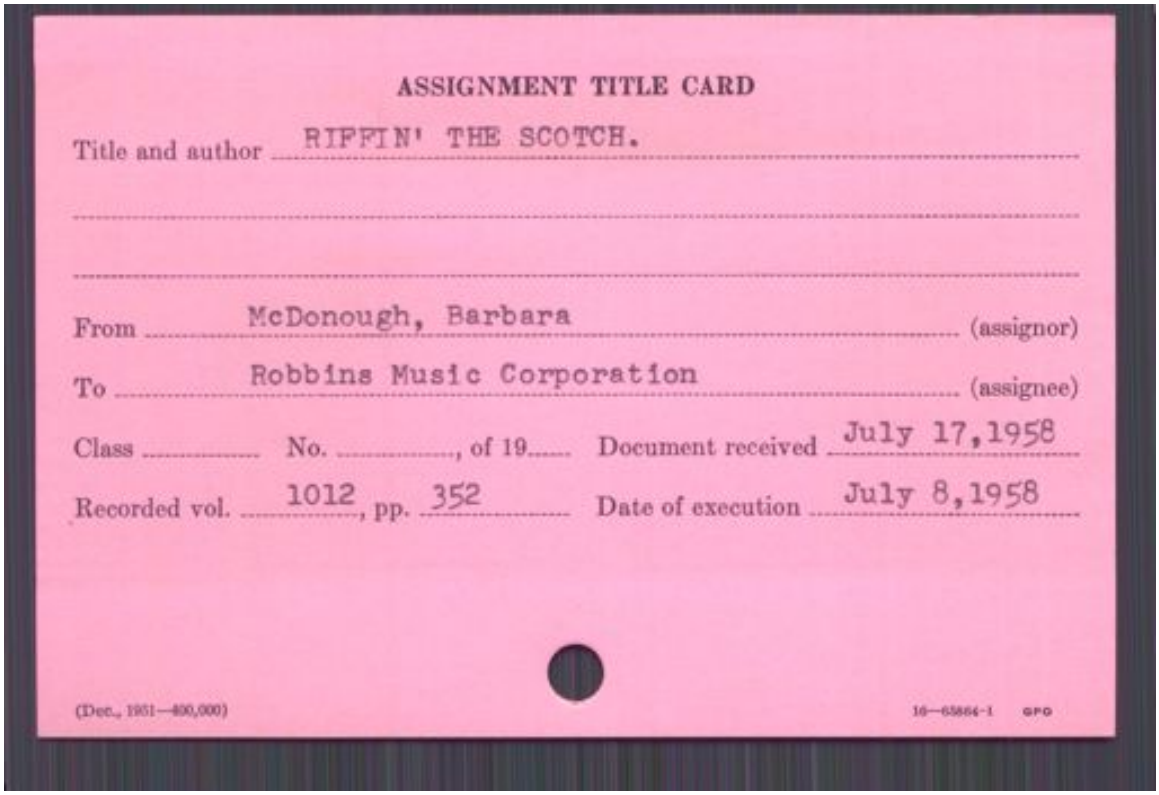
To Robbins Music Corporation (assignee)

Class ..... No. ....., of 19..... Document received July 17, 1958

Recorded vol. 1012, pp. 353 Date of execution July 8, 1958

(Dec., 1951—400,000) 16—65804-1 GPO

Appendix 3



ASSIGNMENT TITLE CARD

Title and author RIFFIN' THE SCOTCH.

From McDonough, Barbara (assignor)

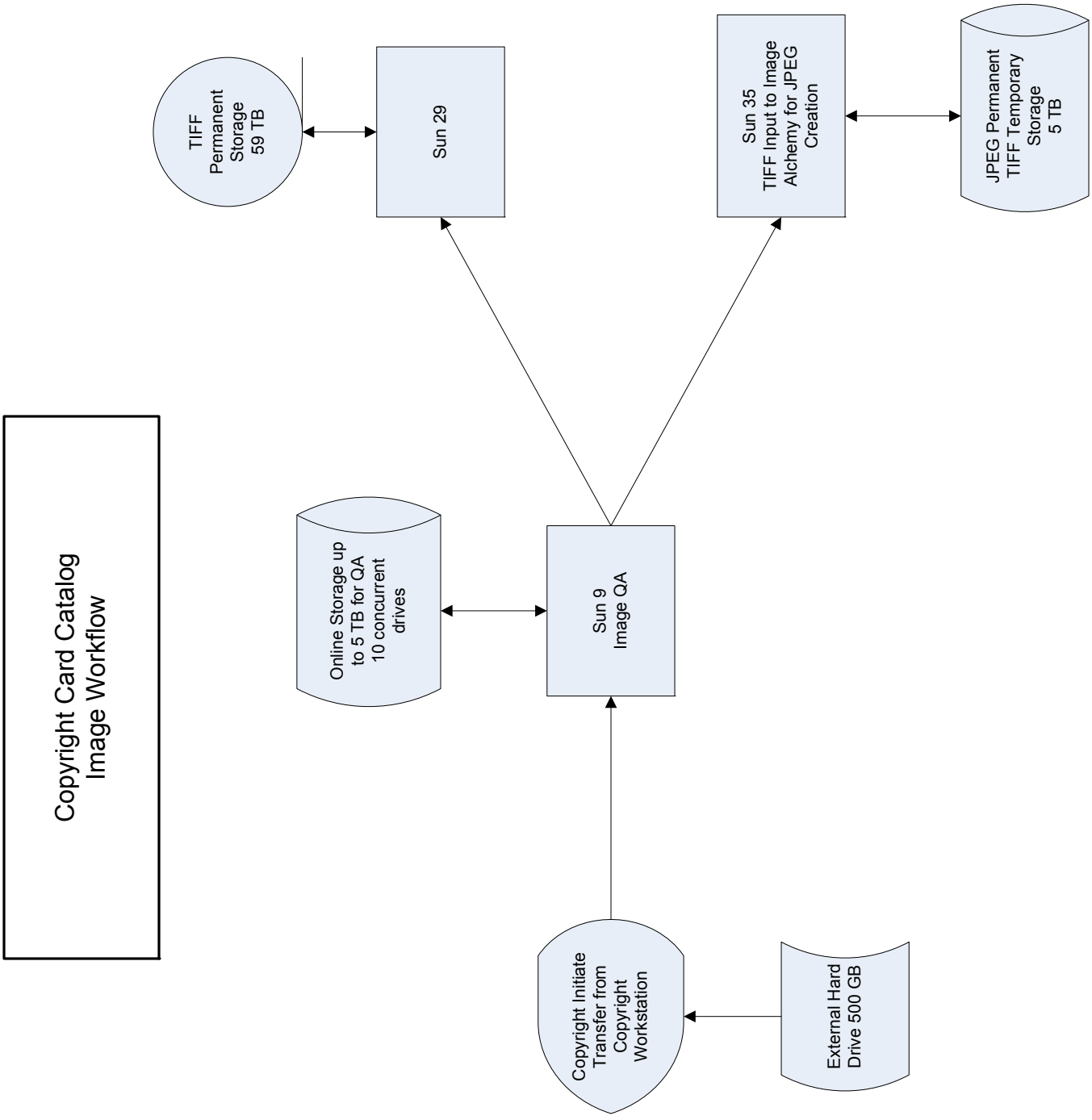
To Robbins Music Corporation (assignee)

Class ..... No. ....., of 19..... Document received July 17,1958

Recorded vol. 1012, pp. 352 Date of execution July 8,1958

(Dec., 1951—400,000) 16—65864-1 GPO

Appendix 4


Copyright Card Catalog
Image Workflow
External Hard
Drive 500 GB
Copyright Initiate
Transfer from
Copyright
Workstation
Online Storage up
to 5 TB for QA
10 concurrent
drives
Sun 9
Image QA
TIFF
Permanent
Storage
59 TB
Sun 29
Sun 35
TIFF Input to Image
Alchemy for JPEG
Creation
JPEG Permanent
TIFF Temporary
Storage
5 TB

File and Bag Naming Conventions

File Naming Convention:

TTTT{time period}{set name}.{sequence}{side}.{format}

- TTTT is a two or four letter prefix indicating the type of record
  - CC – catalog card
  - CCAT – assignment title card
  - CCAR – assignor card
  - CCAE – assignee card
  - CCAX – combined assignor/assignee card
  - RB – record book
  - MF – microfilm
  - CE – catalog of copyright entries
- Time period is a string of up to 8 numbers representing the year or range of years (e.g., 19711977)
- Set name is the title of the volume or microfilm reel or the label on the front of the drawer (e.g., A-AN)
- Sequence is an integer representing the page in a bound volume, the frame in a reel of microfilm, or a card within a drawer
- Side is a suffix that is present only when the verso of a card is scanned or when multiple cards are found stapled or clipped together (e.g., no suffix for single cards with no information on the verso, or "a, b, c, d,…" when a verso is scanned or when multiple card images are scanned for a single entry with the first card face image file carrying the "a" suffix)
- Format is TIF for ingestion and preservation files and JPG or J2K for access files

Examples:

- CCAT19281977ABRO-ACD
- CC19711977KellmKerwin.00001a.tif
- RB1942Music.00001.tif
- MF1987PA123456PA133455.00001.tif
- CE1965Books.00001.tif

Bag Naming Convention:

Original bags for all record types except those beginning with CC will be given names identical to the names of the files they contain up to but not including the period before the sequence number. A similar rule applies for the record types beginning with CC except that the CC is not included in the bag name.

Examples:

- AT19281977ABRO-ACD
- 19711977KellmKerwin
- RB1942Music
- MF1987PA123456PA133455
- CE1965Books

If images in a bag need to be replaced for any reason or if new cards need to be inserted in a bag, then the entire set of image files in the original bag must be rebagged and must carry the same name as the original bag. Images will be ingested and replaced only at the bag level.

COPYRIGHT RECORDS DIGITIZATION PROJECT
Double Blind Data Capture Tool

Logical Record Structures

Control Record

Definition: A record of a bag of images and related metadata that shows the bag's current state and status and the user ID and related information for each of the 3 passes: first data capture, second data capture, and arbitration (This record is necessary because each bag will have its own Access database)

Bag ID	Date bag created	Bag size in number of images	Current pass	Current state	First pass data	Second pass data	Arbitration data
	yyyy-mm-dd		1, 2, A	Ready	User ID	User ID	User ID
				In-process	Date started	Date started	Date started
				Suspended	Date Completed	Date Completed	Date Completed
				Completed	Number of records not selected	Number of records not selected	

User Record

Definition: A record for each authorized user that shows the current bag assigned and the last record processed

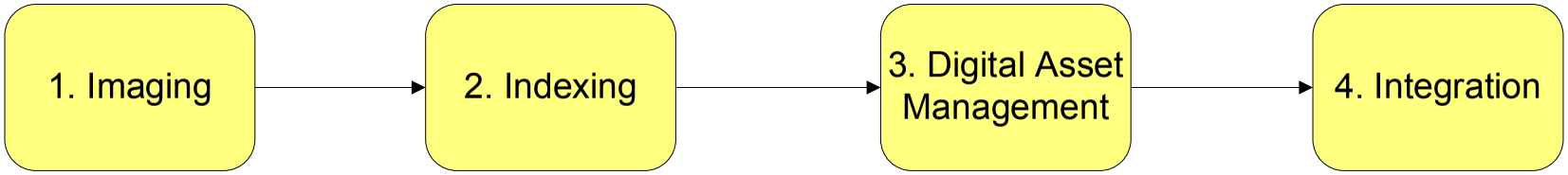
User ID	Bag ID	Last record processed	

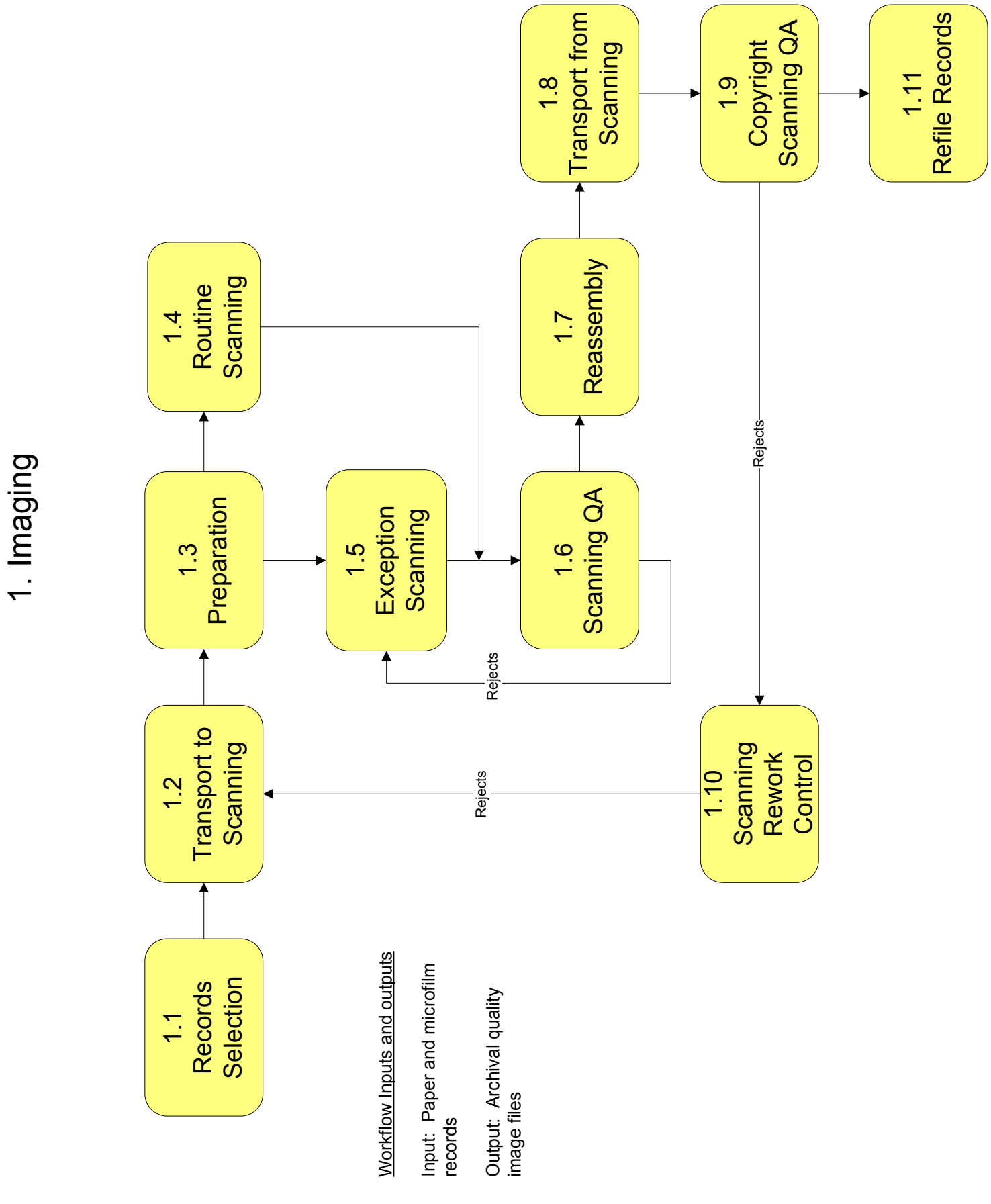
Data Record

Definition: A logical record in the Access database that shows the Bag ID, the current User ID, the data fields from the first pass and the second pass, and the arbitration data fields

Bag ID	Link to image file	First pass fields	Second pass fields	Arbitration pass fields
		Title	Title	Pass fields selected (1 or 2)
		Assignor	Assignor	Pass fields modified (y/n)
		Assignee	Assignee	Selected volume number
		Document title	Document title	Selected page number
		Cross reference	Cross reference	
		Date of recordation	Date of recordation	
		Date of execution	Date of execution	
		Notes (y/n)	Notes (y/n)	
		Bad image (y/n)	Bad image (y/n)	

Copyright Records Digitization Project Workflows



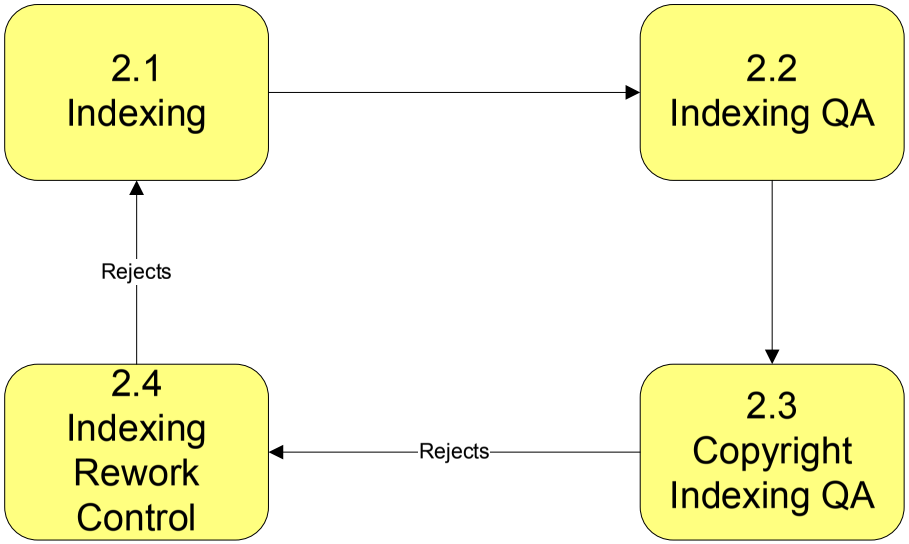

1. Imaging
1.1 Records Selection
1.2 Transport to Scanning
1.3 Preparation
1.4 Routine Scanning
1.5 Exception Scanning
1.6 Scanning QA
1.7 Reassembly
1.8 Transport from Scanning
1.9 Copyright Scanning QA
1.10 Scanning Rework Control
1.11 Refile Records
Rejects
Rejects
Rejects
Workflow Inputs and outputs
Input: Paper and microfilm records
Output: Archival quality image files

2. Indexing

Workflow Inputs and outputs

Input: Compressed derivative image files

Output: Index terms in XML format

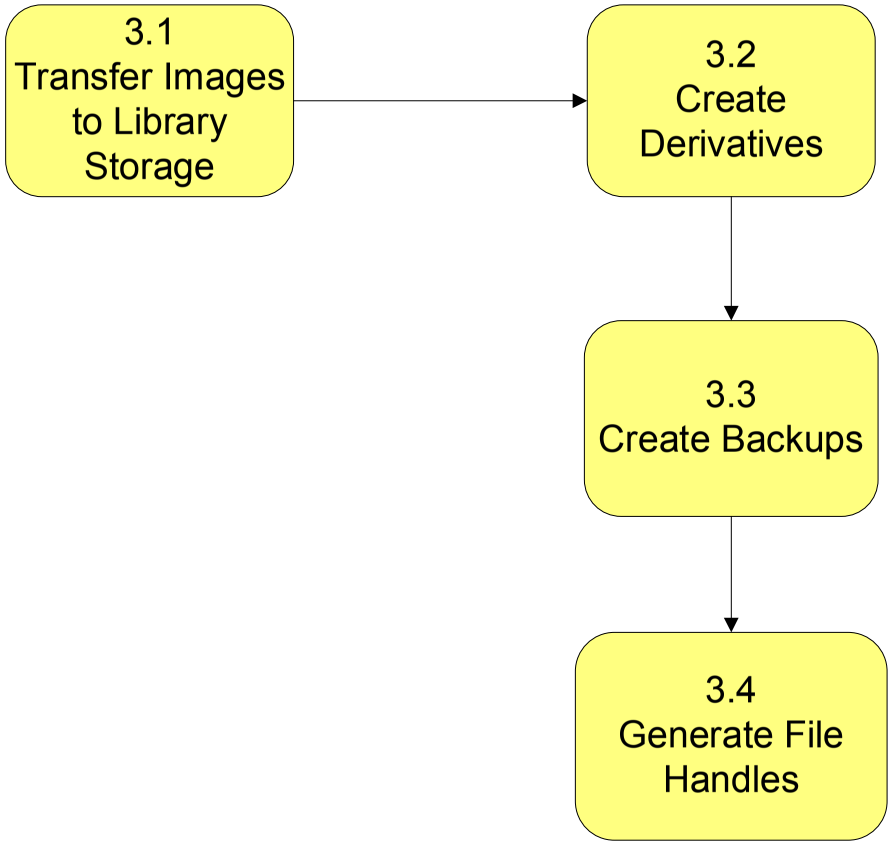


3. Digital Asset Management

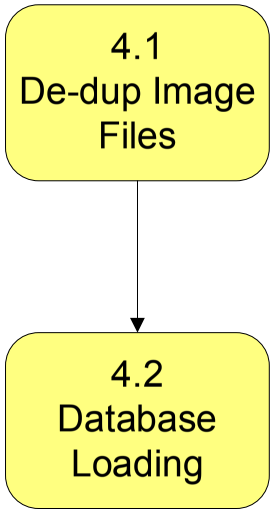
Workflow Inputs and outputs

Input: Archival quality image files

Output: Derivative access image files and file back-ups



4. Integration



Workflow Inputs and outputs

Input: Index terms in XML format and image file handles

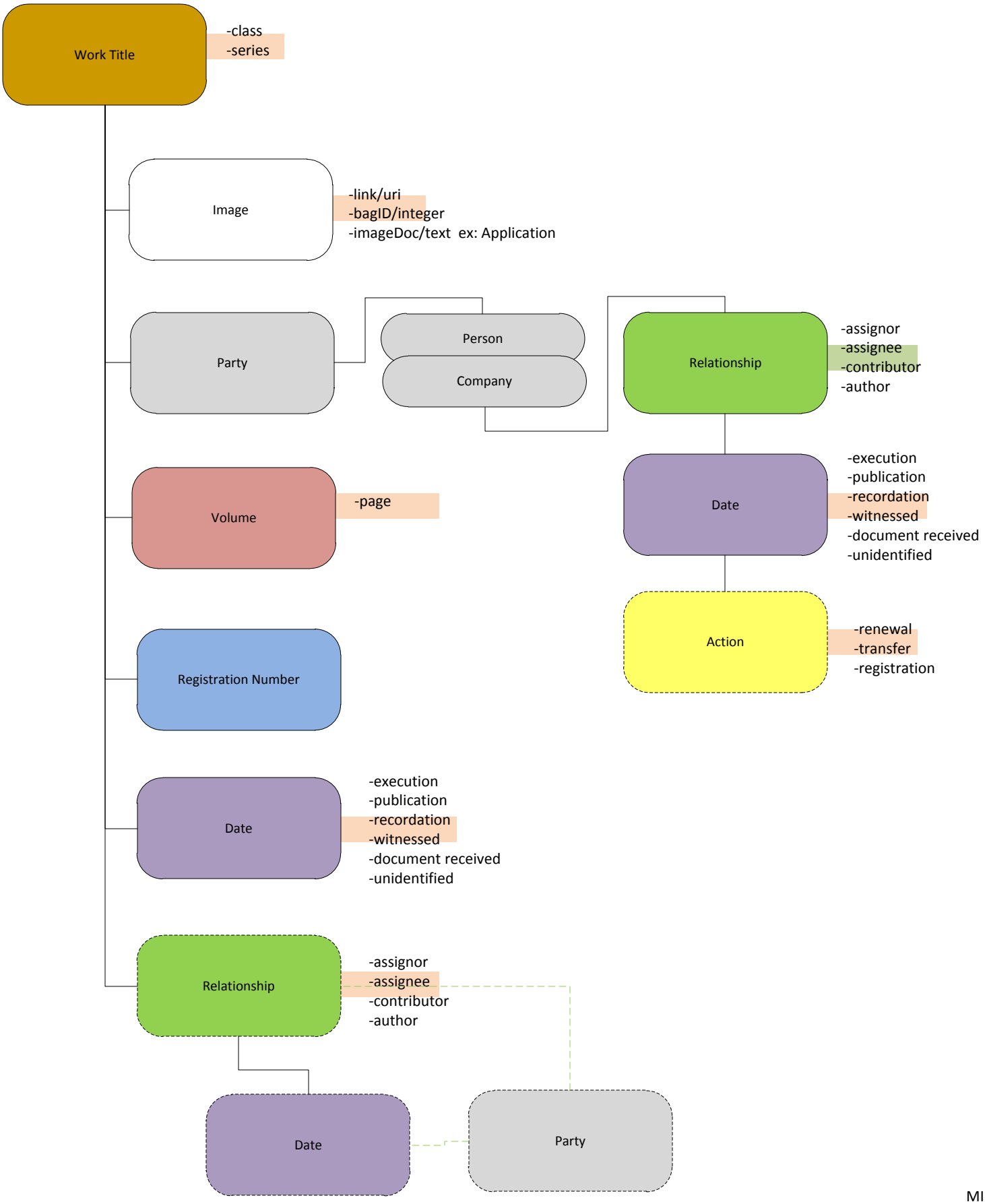
Output: Database records

Appendix 8

XML IN DIAGRAM
VERSION 1

Digital Data Record

DIAGRAM OF XML STRUCTURE



MILLER

METADATA AND XML IN DIAGRAM
VERSION 2

METADATA

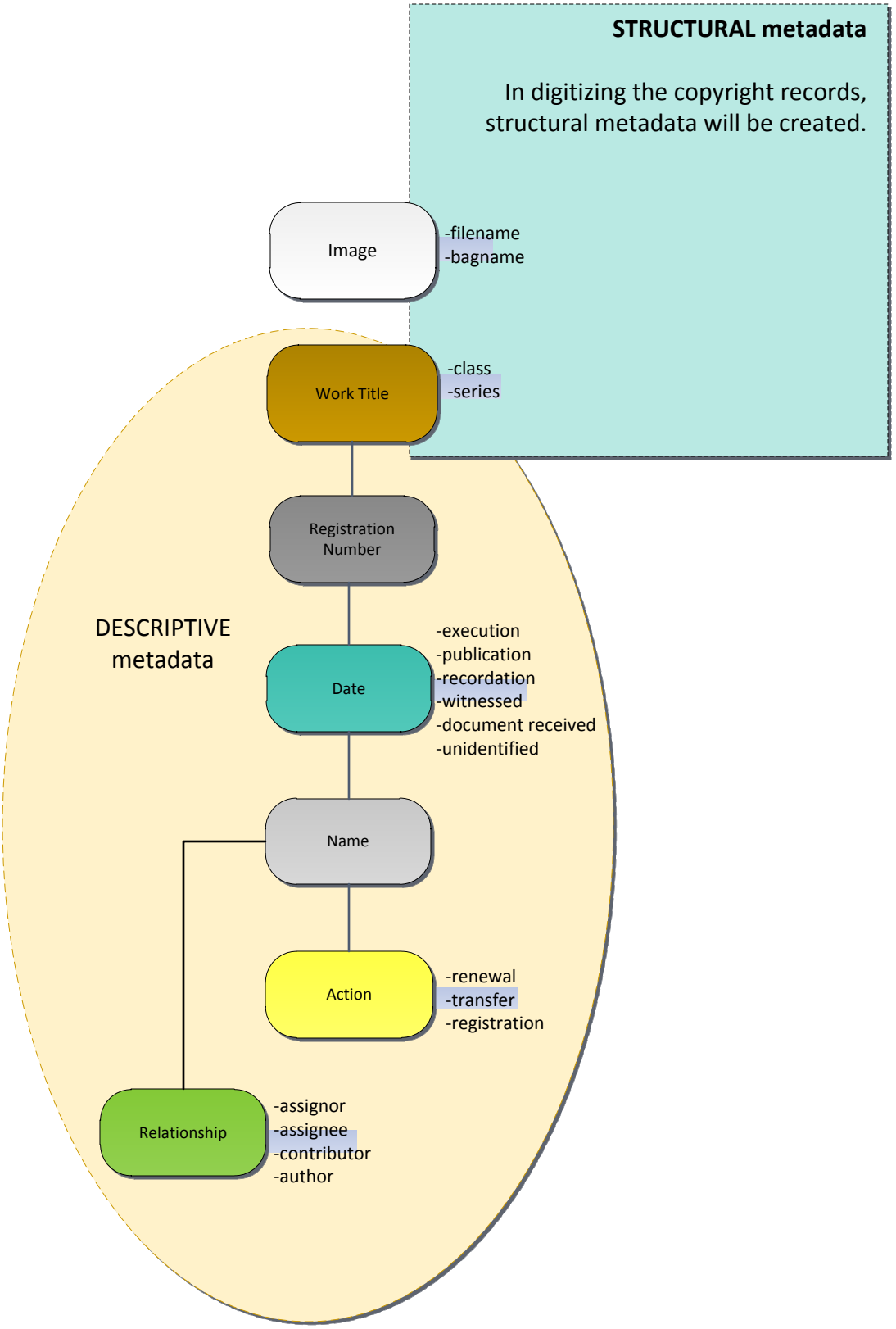
Administrative metadata includes information about ownership, custodianship, access rights and management.

Collectively, these records constitute the **ADMINISTRATIVE metadata** for a work.

A comprehensive collection of this metadata illustrating the copyright history of works is not held by any organization other than the U.S. Office of Copyright within the Library of Congress.

The information used for sorting searching and accessing these records, such as date authorship, and title fit into the category of **DESCRIPTIVE metadata**.

Any metadata used in the ongoing preservation of the digital records can be categorized as PRESERVATION metadata.



```
<?xml version="1.0" encoding="UTF-8"?>
<grammar xmlns="http://relaxng.org/ns/structure/1.0"
    datatypeLibrary="http://www.w3.org/2001/XMLSchema-datatypes">

    <start>
        <element name="collection">
            <oneOrMore>
                <element name="image">
                    <attribute name="filename">
                        <text/>
                    </attribute>
                    <element name="location">
                        <data type="anyURI"/>
                    </element>
                    <interleave>
                        <zeroOrMore>
                            <ref name="element.work_title"/>
                        </zeroOrMore>
                        <zeroOrMore>
                            <element name="registration_number">
                                <data type="integer"/>
                            </element>
                        </zeroOrMore>
                        <zeroOrMore>
                            <ref name="element.name"/>
                            <oneOrMore>
                                <ref name="element.relationship"/>
                            </oneOrMore>
                        </zeroOrMore>

                        <zeroOrMore>
                            <ref name="element.action"/>
                        </zeroOrMore>
                        <oneOrMore>
                            <ref name="element.date"/>
                        </oneOrMore>
                    </interleave>
                </element>
            </oneOrMore>

        </element>
    </start>

    <define name="element.work_title">
        <element name="work_title">

            <zeroOrMore>
```

```
                <attribute name="class">
                    <choice>
                        <value>A</value>
                        <value>B</value>
                        <value>C</value>
                        <value>D</value>
                        <value>E</value>
                        <value>F</value>
                        <value>G</value>
                    </choice>
                </attribute>
            </zeroOrMore>
            <zeroOrMore>
                <attribute name="series">
                    <choice>
                        <value>published</value>
                        <value>unpublished</value>
                        <value>unstated</value>
                    </choice>
                </attribute>
            </zeroOrMore>
            <text/>
        </element>
    </define>

    <define name="element.date">
        <element name="date">
            <attribute name="of">
                <choice>
                    <value>execution</value>
                    <value>publication</value>
                    <value>recordation</value>
                    <value>witnessed</value>
                    <value>document received</value>
                    <value>undefined</value>
                    <value>official signature</value>
                </choice>
            </attribute>
            <data type="date"/>
        </element>
    </define>

    <define name="element.name">
        <element name="name">
            <optional>
                <element name="firstName">
                    <text/>
```

```
            </element>
        </optional>
        <optional>
            <element name="middleName">
                <text/>
            </element>
        </optional>
        <choice>
            <element name="lastName">
                <text/>
            </element>
            <element name="company">
                <text/>
            </element>
        </choice>


        <zeroOrMore>
            <element name="authorityName">
                <text/>
            </element>
        </zeroOrMore>

    </element>
</define>

<define name="element.relationship">
    <element name="relationship">
        <attribute name="relationship_type">
            <choice>
                <value>assignorFrom</value>
                <value>assigneeTo</value>
                <value>contributor</value>
                <value>author</value>
                <value>other</value>
                <value>undefined</value>
            </choice>
        </attribute>
    </element>
</define>

<define name="element.action">
    <element name="action">
        <choice>
            <value>renewal</value>
            <value>transfer or assignment</value>
            <value>registration</value>
```

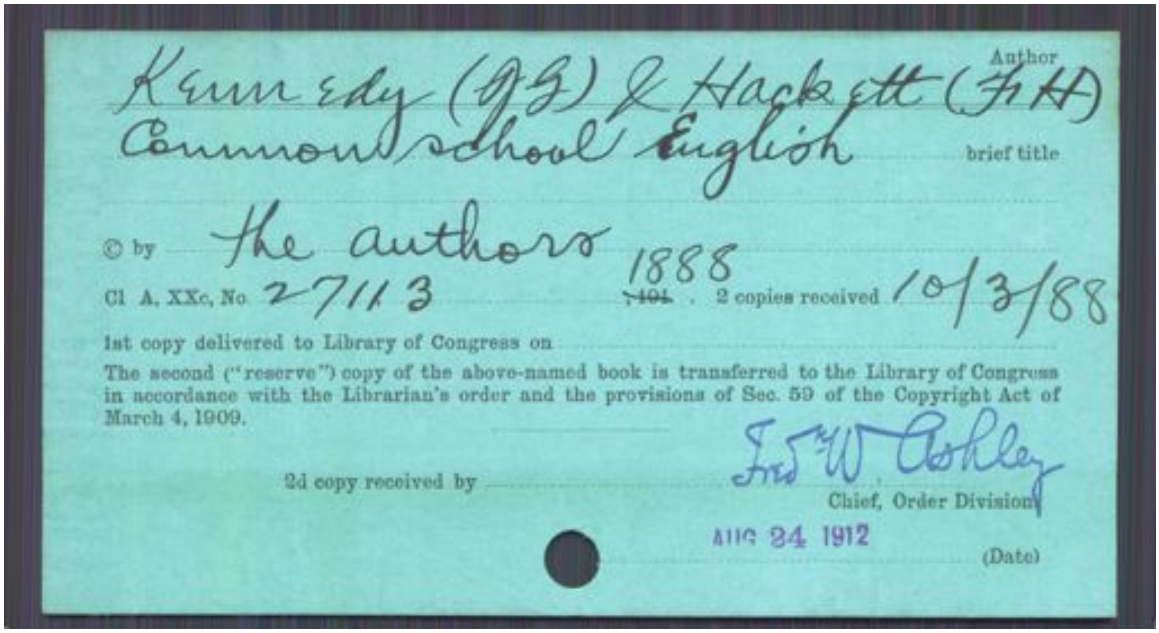
```
            </choice>
        </element>
    </define>


</grammar>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<RecordedDocument>
    <Record>
        <RecordID>
            <Attribute name="type">
                <ValueChoice>
                    <Value>volume</Value>
                    <Value>reel</Value>
                </ValueChoice>
            </Attribute>
<!-- recordID is the combination of the volume or reel no and the page or
frame no ragne
    separted by "/"; example: "1301/77-81." Attribute type makes the
distinction -->
        </RecordID>
    </Record>

    <Work>
        <WorkTitle>Given work title </WorkTitle>
        <WorkClass>If given</WorkClass>
        <RegistrationNo>If given</RegistrationNo>
        <RegistrationYear>If given</RegistrationYear>
    </Work>

    <Recordation>
        <RecordationDate>Dated listed; ex: 1968-03-28 </RecordationDate>
        <AssignorAgent>text</AssignorAgent>
        <AssigneeAgent>text</AssigneeAgent>
    </Recordation>

    <LinkingImageIdentifier>
        <LinkingImageIdentifierType>URL</LinkingImageIdentifierType>
        <LinkingImageIdentifierValue>Persistant  URL
Path</LinkingImageIdentifierValue>
    </LinkingImageIdentifier>
</RecordedDocument>
```

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<?xml-model
  href="file:/C:/Users/Owner/Documents/Library%20Science%20Studies/LIS%20Electronic%20Publishing/Version2_Copy
  type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"
?>
<?xml-model
  href="../../Library%20Science%20Studies/LIS%20Electronic%20Publishing/Version2_CopyrightDDR.rng"
  type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"
?>
<collection>
  <image filename="CC1870189KELLM-KERWIN.00003.300">
    <location>uri.tif</location>
    <work_title class="A" series="unstated">Common School English</work_title>
    <registration_number>27113</registration_number>
    <date of="document received">1888-10-03</date>
    <name>
      <firstName>J</firstName>
      <middleName>G</middleName>
      <lastName>Kennedy</lastName>
    </name>
    <relationship relationship_type="author"/>
    <name>
      <firstName>F</firstName>
      <middleName>H</middleName>
      <lastName>Hackett</lastName>
    </name>
    <relationship relationship_type="author"/>
    <action>registration</action>
  </image>
  <image filename="CC18701941HARR-HIK.00011.300">
    <location>uri.tif</location>
    <work_title>
      Irene, the Stubborn Girl also known as "Ten Eleven Fifth"
    </work_title>
    <name>
      <firstName>Eric</firstName>
      <lastName>Hatch</lastName>
    </name>
    <relationship relationship_type="assignorFrom"/>
    <name>
      <firstName>Paul</firstName>
      <middleName>R</middleName>
      <lastName>Reynolds</lastName>
    </name>
    <relationship relationship_type="author"/>
    <name>
      <company>Universal Pictures Corporation</company>
    </name>
    <relationship relationship_type="assigneeTo"/>
    <date of="document received">1935-11-30</date>
  </image>
</collection>
```

Appendix 10



Kennedy (DJ) & Hackett (FH) Author
Common school English brief title

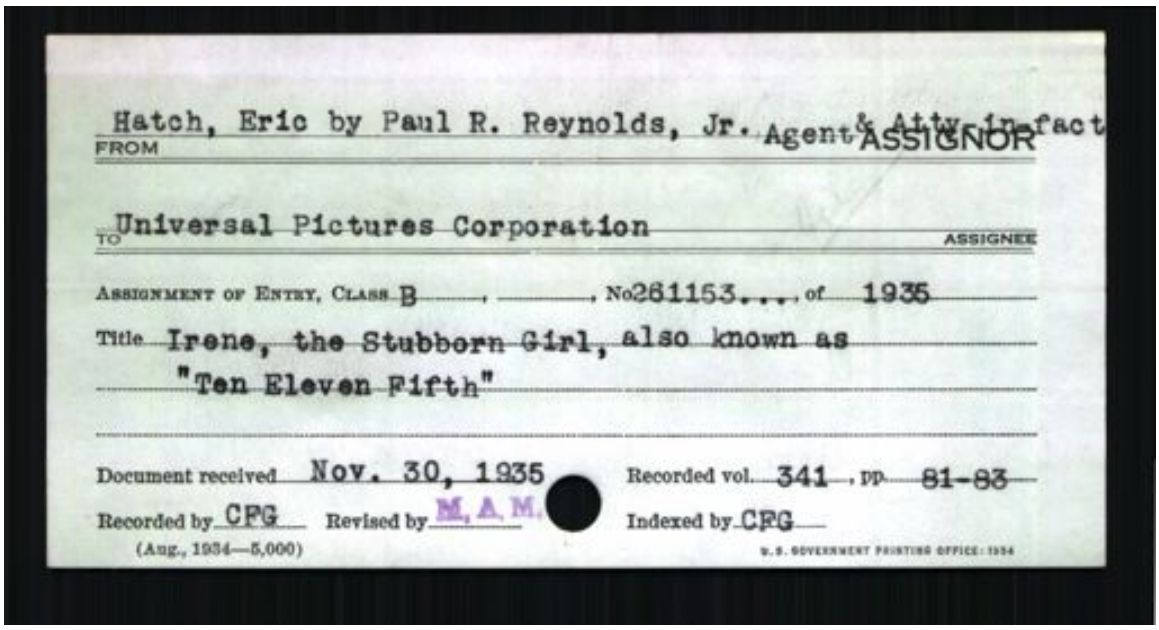
© by the authors 1888
Cl A, XXc, No 27113 ~~491~~ . 2 copies received 10/3/88

1st copy delivered to Library of Congress on

The second ("reserve") copy of the above-named book is transferred to the Library of Congress in accordance with the Librarian's order and the provisions of Sec. 59 of the Copyright Act of March 4, 1909.

2d copy received by Thos W Ashley
Chief, Order Division

AUG 24 1912 (Date)



Hatch, Eric by Paul R. Reynolds, Jr. Agent & Atty in fact
FROM ASSIGNOR

Universal Pictures Corporation
TO ASSIGNEE

ASSIGNMENT OF ENTRY, CLASS B, , No261153..., of 1935

Title Irene, the Stubborn Girl, also known as "Ten Eleven Fifth"

Document received Nov. 30, 1935 Recorded vol. 341, pp 81-83

Recorded by CFG Revised by M.A.M. Indexed by CFG

(Aug., 1934—5,000) U.S. GOVERNMENT PRINTING OFFICE: 1934