Representing Archival Descriptive Metadata in a DSpace Environment

Patricia Galloway
School of Information, UT-Austin
I University Station D7000
Austin, TX 78712
512-232-9220
galloway@ischool.utexas.edu

ABSTRACT

Archiving born-digital materials brings with it the problem of how—or if—to implement standard hierarchical modes of archival description in an environment where the user can potentially make granular searches of repository item-level metadata and the primary materials themselves, thus bypassing mediating finding aids and even disregarding collection and series boundaries altogether. In this essay I will discuss some of the problems of providing archival description as metadata in the case of digital archival materials, especially as seen in the context of the DSpace digital repository software environment, which instantiates a specific information structure itself. Examples will be taken from practice as developed in work at the School of Information on collections ingested into the School's institutional repository, and some additional unconventional descriptive practices will be described.

Categories and Subject Descriptors

H.3.7 [**Information Systems**]: Information Storage and Retrieval: Digital Libraries – *collection, standards, systems issues.*

General Terms

Design, Standardization, Theory

Keywords

Archival description, collection-level, DSpace, Dublin Core, item-level, metadata.

1. INTRODUCTION

As Elizabeth Yakel has recently articulated the issue, archival description is a representational practice, developed historically to support a specific and still culturally circumscribed concept of information creation and aggregation.[14] Like other cataloging practices it entails restriction of the view of archival materials to that of the describing archivist, both personal and as assimilated from institutional standards. In the case of paper collections, there is a real reason for succinct descriptions to provide the user with a way of making quick gross distinctions between what she wants to see and doesn't want to see: it is impractical to read through even a fraction of available paper collections to discover whether they are relevant, so so-called "finding aid" descriptions have come to seem, at least for users like archivists and historians trained to see them that way, a natural proxy for the collections themselves. As long as the describing archivist and the user share world view, vocabulary, and semantics, traditional archival description seems to work fairly well, although the limited track record for current more codified (and supposedly improved) practice and the lack of user studies for archival description in general must make this a guarded observation.

The introduction of computers to aid archival work, first seen as a convenient way to computerize finding aids, had the effect of pushing descriptive practice toward uniformity as the advantages of centralized pools of information about archival resources began to be appreciated.[9] But there is a big difference between automated access to information about archival records when the records themselves are paper, and the same access when the records are born-digital and potentially equally accessible, capable of competing with archival description to define themselves. The advent of digital archival repositories and the broad global access they permit opens up entirely new problems of representation for archivists. Discussions of the possibilities have ranged from the suggestion that archival intervention in system design at the point of creation of digital objects may provide an abundance of metadata to be repurposed to fulfill archival functions; to the suggestion that in the new world of metadata-accompanied digital objects, archival description may no longer be needed at all.[3, 5]

Yet I think there will be continuing value in traditional aggregational finding aids if they are seen as situated expert commentary on archival collections that represent the related production of an individual, group, or institution. Such descriptions convey curatorial authority and provide information actually not likely to be generated through the normal creation and use of digital objects (as, for example, a substantial biographical note specifying information about the creator's specific relationship to the collection in question). There is also value in multiple virtual archival "arrangements" of digital materials that express a view of the logical relationships among them while allowing the preservation of the creator's own virtual arrangement alongside that of the archivist. In short, it makes sense (not just to avoid disturbing established practice) to attempt to incorporate conventional archival "arrangement" and description into the work of a digital archival repository, if not as the sole authoritative last word on the subject, then as a value-added and evolving tool for the user alongside other virtual arrangements and descriptions supplied by both published academic researchers and, say, groups of lay researchers like genealogists or simply individuals whose "my collection" arrangement/description might be contributed to the repository in some way. New thinking by historians and other researchers interested in born-digital materials is nurturing demands for other ways to create views of digital resources, including sophisticated data mining heuristics and visualization tools.[1]

2. DSPACE REPRESENTATION

To save finding aids alongside their born-digital referents, digital repositories need to provide support for such aggregate descriptive metadata. What few have discussed at any length is the significance of the metadata infrastructure and the access interface of the digital repository itself as a representational practice that will affect the way archival materials are described. The DSpace digital repository software was designed by MIT librarians and Hewlett-Packard programmers to accommodate the self-archiving of digital objects owned and controlled by MIT faculty belonging to different administrative units and communities of practice: departments, schools, and institutes. Another assumption was that self-archiving through these units would be preceded by professional library construction of the fonds-like groupings into which these objects would go and followed by professional catalogers' "grooming" of the creator-supplied metadata. At the same time, there was a serious concern to comply with and implement to some degree the basic functionality of the widely-accepted Open Archival Information System (OAIS) model for an archival digital repository, as it was assumed that the materials held in a DSpace instance would be held for a long time or permanently.[8]. For these reasons DSpace actually accommodates a fair amount of metadata that applies to all of the functions of a digital archives, but with an interesting mixture of automated harvesting and artisinal supply of cataloging metadata at the individual object level. As the user interface stands now, however, the opportunity to supply many of the metadata elements that are implemented in the standard release database structure is not available through the individual submission interface, although these elements can be exploited via after-the-fact editing of item-level metadata through the collection administrator interface or uploaded directly through the batch ingest process (or, of course, the submission user interface can be rewritten).

The overall representation of information in a DSpace instance is hierarchical because of MIT's original vision of a single repository operated by information professionals and hosting materials that had to be segregated conceptually into heterogeneous series, primarily conceived as "publications" in the sense of making them public to the world. Originally this structure consisted of Communities, Collections within those communities, and Items within collections. Shortly it became obvious that an intermediate level was needed between communities and collections, so the Subcommunity construct was added (identical to Community except that in the standard web interface one descriptive field created during setup is actually displayed, whereas for the Community it is not—see below). Subcommunities, nested within communities, may also have additional subcommunities nested within them, and so on: in theory subcommunities can be infinitely nested, but in practice users are advised to aim for as flat a structure as can accommodate their materials.[6] Collections as aggregations of items must be placed within communities or subcommunities, and any number of them may be created.

From the standpoint of the operator of a DSpace instance, most of its functions are variably accessible (intentionally so) through a web-based user interface, depending on the permissions assigned to the user ("e-person") or group of users. The permission system is quite elaborate, allowing for the restriction of access to an individual object to a single person if necessary, but in practice there are several authentication sets, manifested as roles, that apply broadly to the different DSpace constructs. The *Administrator* has all permissions for every action available through any part of the web user interface: only the Administrator may set up the community–subcommunity–collection structure and add descriptive information to the community and subcommunity structures. Initially the Administrator was also responsible for descriptive information for the collections, but early on it was decided to create an intermediary role, *Collection Administrator*, so as to delegate some of the administration of collections: the collection administrator cannot create communities, subcommunities, or collections, but s/he does have some editing capabilities (adding information about the collection, for example) and can manage collections by assigning to individual e-people or groups of e-people permissions to perform various workflow actions on items, like submitting items and metadata about them, editing, and approving submissions. Once assigned permission to perform these workflow tasks, *Submitters* have the ability to place items and metadata about them into already-created collections within the repository, but not to edit them after submission is complete.

Articulated in this way, the standard DSpace distribution has some potential to accommodate traditional hierarchical representation of digital collections (and its roles even interestingly reflect some of the division of labor in the enterprise of archival arrangement and description). As an archival educator I find that it makes a robust environment in which aspiring digital archivists can grapple with many of the problems of digital archiving while at the same time being forced to become mindful users of a specific representation system to archive and describe useful collections of digital materials. Because DSpace is also informed by new ideas from the digital libraries arena, it also provides some capabilities that call archival tradition into question. It is, for example, possible for an item to be listed in more than one collection and it is possible for a collection to be created that will "contain" only items already present in other collections, thus allowing for an entirely virtual rearrangement and redescription to be created. Because the DSpace software is open-source and because it sits on top of an SQL database that contains all the system's metadata (descriptive, technical, structural, administrative) and a file system into which items are put according not to any kind of physical concerns but rather to storage efficiency, it is in principle possible for anyone to add to a DSpace instance the ability to expose any amount of metadata to the user in a variety of ways and indeed to treat the entire contents of the repository as raw data to be mined, harvested, and repurposed. This means that as new capabilities are added to DSpace through the open-source process, it is possible for users of the software to continue to contemplate new approaches and new potential features.

Within the structure that can be created using the DSpace-defined objects Community, Subcommunity, Collection, Item, and Bundle, there is a broad range of metadata called for to describe a fully-developed hierarchical structure, not all of which is directly available to the user of collections. It should be noted that within DSpace, and from a data point of view, all objects—e-people, groups, communities, subcommunities, collections, items, bundles, and bitstreams—are treated the same: they are objects with unique identifiers that belong to a single sequence, although each category has a different set of metadata associated with it. Note that these are the core defined objects available within the system that cannot be altered without additional programming. In principle all the metadata elements can be made available to any user if additional parts are added to the user interface. At present the descriptive elements that belong to communities, subcommunities, and collections are displayed authoritatively on the respective "container's" homepage, whereas metadata elements that apply to the item level (more conventionally the content) are displayed on familiar library OPAC-style short and longer bibliographic displays. Metadata for e-people, groups, bundles, and bitstreams are much more restricted from view, since they mostly pertain to the control of access. Except for a few elements automatically populated during the ingest process, metadata associated with the DSpace History module, which records an audit trail of all changes that are made to any given object within the system, is not presently available at all through the web user interface or indeed to any e-person role defined within it except in item-editing mode.

3. MAPPING DSPACE TO DACS[1]

I will list these metadata elements here with some explanations and a rough mapping to the archival DACS standard[2] so that in the discussions of cases that follow, it will become clear how we have mapped a more conventional archival descriptive practice onto the DSpace structure.

Community and Subcommunity (can only be created by the overall Administrator. Note that the creation of these structural "container" objects amounts to an act of structuration analogous to arrangement and specifically allocated shelving).
Name [DACS: Title]
Short Description (for Community, not displayed; for Subcommunity, displayed as part of a list of subcommunities on Community page)
Introductory text [DACS: Administrative/Biographical History, Scope and Content, Custodial History]
Copyright text [DACS: Conditions Governing Reproduction and Use]
Sidebar text
Logo (Displayed on Subcommunity page)
Authorizations [DACS: Conditions Governing Access, Physical Access, Technical Access]

Collection (can only be created by Administrator; can be edited by Collection Administrator once created)
Name [DACS: Title]
Short description (displayed for the Collection listing on the parent Community or Subcommunity page)
Introductory text (Displayed on the Collection page) [DACS: Administrative/Biographical History, Scope and Content, Custodial History]
Copyright text [DACS: Conditions Governing Reproduction and Use]
Sidebar text
License [DACS: Conditions Governing Reproduction and Use]
Provenance [DACS: Acquisition and Appraisal Elements]
Logo (Displayed on the Collection page)
(information that follows sets up permissions and process definitions that enable the ingest of digital objects into the collection)
Submission workflow
    Submitters
    Accept/Reject step
    Accept/Reject/Edit Metadata step
    Edit Metadata step
Collection administrators
Item template
Authorizations [DACS: Conditions Governing Access, Physical Access, Technical Access]

Item (the item is the referent for the Qualified Dublin Core [QDC] metadata supplied with standard DSpace. In principle, other relevant metadata sets to support specific formats or preservation procedures for digital objects can be added to the supplied QDC, whose most significant use is for resource discovery.) [DACS: Reference Code, Title, Date, Extent, Name of Creator(s), plus Related Materials Elements]

Bundle ("A Bundle represents all of the related Bitstreams that are required to render a manifestation of an Item.")[7]
Authorizations [DACS: Conditions Governing Access, Physical Access, Technical Access]

Bitstream
Name (every bitstream bundle consists of a group of license bitstreams as well as any ingested content bitstream(s))
Source (harvested automatically)
Description (normally not present unless specifically added; not displayed) [note here we used "access copy" and "archival copy" where relevant, and this displays with the bitstream on the item page]
Format (code, assigned automatically; set to 1 or Unknown if a user value is added)
User Format Description (when DSpace does not recognize format or does not have details, user may add this information) [DACS: Languages and Scripts of the Material?]

The DSpace system thus provides a specific representation of its contents. Although DSpace is open-source and can be altered, there are many ways in which its fundamental assumptions, at least as so far implemented, imply as hierarchical a representational system as is envisaged in archival practice, especially since the addition of arbitrarily nested subcommunities. But it takes some careful thought to achieve a hierarchical representation, and doing so breaks the DSpace assumption that it will contain synthetic collections. Hence some of the standard methods, especially

[1] Describing Archives: a Content Standard (DACS) represents the current recommended standard for American archival descriptive practice; it is based on and close to the General International Standard Archival Description (ISAD[G]) and is easily accommodated in Encoded Archival Description (EAD) markup.

of search within the repository, meant to be applied to defined objects in DSpace are not appropriate to the mappings suggested here. In addition to the representational constraints, DSpace limits the creation of information structures to e-people filling a few roles, in practice usually allotted not to ordinary submitters or users but to librarians and other information professionals. Control over other actions has been left outside the web user interface entirely, handing it to the DSpace system administrator who has privileged access to the configuration and operation of DSpace itself from the operating system command line. As DSpace has evolved so far, this is probably a good thing, as even those professionals who use it sometimes find it difficult to understand in detail the interaction of parts that creates the functionality of the authorization system, which is the basis for DSpace security.

4. MAPPING ARCHIVAL REPRESENTATION ONTO DSPACE

Thus the DSpace support for collection representation is not a precise fit for archival practice if used as intended, but it does (or can) cover a large amount of the recommended DACS descriptive elements by placing them in the unspecified "description" fields for the Community, Subcommunity, and Collection "containers." In our work with DSpace it should be said that the students who have created the arrangements and descriptions for materials archived on our DSpace server have not all been archives students: each team working on a collection project has included at least one archives student, generally one who has already been exposed to conventional archival descriptive practice, together with 2-3 students from a concentration in preservation administration, librarianship, digital libraries, or information architecture. This mixture of competencies from communities of practice that neighbor fairly closely in our institution, all having been exposed to a uniform set of core courses on information organization, user issues, research methods, and management, has led to a rich cross-fertilization of practice, supported by a readings- and lecture-based introduction to the issues of digital archives (as distinct from digital libraries, but not so distinct: we see digital libraries as needing to have a digital archives component, especially as they increasingly harbor born-digital objects that they need to preserve). In this context the DSpace repository with its specific functionality becomes a boundary object for students from different backgrounds, and the requirement to complete archiving projects including novel formats and complex intellectual property regimes adds the problem-solving impetus that creates new emergent practices. As a result, we have experimented with different approaches to arrangement and description in solving problems of representation using the native DSpace structuration tools.

4.1.1 General workflow procedures for pre-ingest

The archival received wisdom about the dialectic between processing and description seems to hold just as true for digital objects as for papers. The canonical sequence of steps in arrangement and description, as represented in the current Society of American Archivists (SAA) basic manual, are as follows:

- Accessioning archival records
- Establishing contextual information for arrangement and description
- Arranging the records
- Physically processing the records
- Describing the records
- Developing access tools[11]

This specific sequence is far more appropriate to materials whose physical manifestation may be considered to be fairly inert than it is to digital objects. In practice we have found that our first task must be to inventory the materials to be ingested and ideally establish the equivalent of a conservation condition report. The OAIS reference model assumes a dynamic situation where the creator alone or working with a digital archivist crafts a specific agreement with the repository about transfer that specifies a great deal about the nature of the records. In the projects we have carried out, we have found that few creators have a detailed idea of their digital holdings, with the exception of more recent materials in active use. Further, like most digital repository operators, we have found that the ideal of setting up an ongoing relationship with records creators, wherein they make deposits to the repository themselves, may some day come to pass, but for the moment we are confronting two general situations: either, as in the case of the Harry Ransom Center's materials discussed below, we are working with collections already accessioned (and indeed where the creator may be dead) but where the digital component of the collection is unprocessed (and may even have been ignored); or, as in the case of materials belonging to the School of Information itself, we are catching up to capturing materials that have been partly forgotten and are in danger of loss. In both cases inventory, condition report, and formal agreement have to be created.

Because of the nature of digital materials, it is perfectly possible to inventory them from a distance if proper access has been arranged or to do so from media already in custody, but the condition report must enable further stages of processing by discovering whether files are readable and if so how—hence one can rarely inventory without processing to some degree. We would like to think that the so-called digital archaeology processes of recovering outdated file formats is something that might in a better-informed future be avoidable, but we fear that the continued constant evolution of both proprietary systems and technological possibilities will ensure that such will not be the case any time soon, if ever. These procedures, however, are necessary for the recovery of the technical metadata relating to file creation and rendering that are vital to long-term preservation, and we have adopted a practice of recovering said metadata into spreadsheets representing standard schemes, using metadata-harvesting software where possible.

As the SAA arrangement and description handbook suggests, arrangement can be seen as a fundamental activity of description. If anything the deciphering of the original order maintained by the user of digital objects is more significant to understanding the "information ecology" that they represent for the creator/user than is the case with paper because idiosyncratic directory structures and file naming conventions can be such a barrier to access, even to the creator himself. But the "original order" arrangement of digital objects is a fluid thing, much influenced by the interface through which that arrangement is understood by the owner of the objects, who may think in terms of icons on a desktop, a dynamically-sortable flat list of files, or a detailed hierarchical structure—or all of these at different times. Further, as processors of paper archives well know, any "original order" is usually a snapshot, a state of the materials as and when acquired, and in the case of digital materials it is not at all clear exactly what "original order" should be preserved or restored. Digital collections begin to show us that many such orders are both discoverable within the materials and their system context and preservable in such a way as to make it possible to view a prior state of their relations, much as sophisticated transactional databases with full audit trails can be "rolled back" to a previous state if needed. Capturing what the user sees and uses as well as the "underlying" structure provided by the system in some way is difficult though not impossible, and we do so by recording directory structures, evidence of versioning, and file modification date stamps for the sake of a possible need for this information. Preserving and representing such structures are presently usually carried out as secondary practices, based on archived records of the ordered relationships among the bitstreams archived, although we are now experimenting with representing directory structures directly in DSpace (see below).

4.1.2 Representation according to local archival practice: HRC Materials

Before digital objects can be ingested into DSpace, they must have a place to go, so to a large degree the upper reaches of the archival hierarchy (fonds, series) must be constructed in advance and derived from the inventory—as well as from institutional practice. In what follows I will use a case study based on archival practice as applied to the creation of DSpace collections for an established collecting archives in order to demonstrate how existing archival descriptive practice interacts with the systemic support provided by one archival software system. For contrast I will also draw upon less constrained cases from our own institutional collections.

Since 2005 we have been archiving collections of digital materials held by the Harry Ransom Humanities Research Center (HRC) at the University of Texas-Austin. The HRC, founded in 1957, is a collecting archives based on rare book and manuscript collections acquired by the University of Texas library since the nineteenth century but developed since 1958 to focus on literature and the humanities and to acquire manuscript materials rather than rare books. Because of its origins, however, until 1990 HRC librarians cataloged manuscript collections at item level into a card catalog. Since 1990 HRC archivists have developed finding aids according to an archival model, such that HRC finding aids now conventionally contain:

Biographical Sketch
Scope and Content Note *
Series Descriptions
Folder List *
Correspondents List[2]

Completed finding aids are available on the HRC website and through the Texas Archival Resources Online (TARO) website, for both of which they have been encoded in EAD. When we began working with digital materials it was clear that HRC practices did not actually envisage coping with digital materials in any other way than exactly as paper and artifact materials were treated. Fortunately, the still-lingering tradition of manuscript cataloging at the HRC (the pre-1990 card catalog of manuscripts is still actively in use) at least made item-level metadata seem a familiar and valued concept. The first collection we worked with, however, was also the HRC's first predominantly-digital collection: that of hypertext novelist Michael Joyce. Working with this collection was especially interesting from a description point of view because since the collection was new, there was no existing finding aid: the structure of the collection might be based significantly on the digital materials in it.

Ironically, however, considering the fact that a good part of Michael Joyce's significance as an author rests on hypertext novels that can only be written and perceived through the mediation of a computer, the digital structure did not dominate. Because Joyce had frequently printed his digital files and had written revisions on the printouts before making digital revisions; because the major interest of HRC is the creative process, and hence the preservation of all versions of a work including in this case the paper ones; and because the arrangement of paper records is HRC's most familiar mode and the paper versions had been accessioned and arranged first—the *paper* versions of Joyce's digital files took precedence in the arrangement process and their arrangement became the structure that the digital collection mirrored.[13] Now granted, the paper versions of digital files often had holograph revisions on them and therefore contained more net creative content than the digital versions from which they derived (the metadata quantity, however, ran in the opposite direction: the paper versions often lacked any means of dating, for example, such that the revision history of a given work could only be certainly worked out with reference to the sequent digital files). Hence the paper arrangement, anchored most notably to specific works and to the author's productive roles, was held in some sense to represent the "original order" of the collection and governed the structure of the digital Joyce collection, even though the digital materials themselves were found grouped chronologically as often as functionally. Then, as Stollar Peters observed, the archival fonds, equated to the whole of the collection obtained from Michael Joyce, was mapped onto a DSpace (sub)community. This subcommunity contained series, which were themselves instantiated as subcommunities, and the series in turn might contain collections or subseries, which in turn might be configured as DSpace subcommunities containing collections or might simply be configured as DSpace collections:[3]

- ❖ Harry Ransom Humanities Research Center, Michael Joyce Project (subcommunity of "Special Projects" community in the School of Information repository)
    - ➢ Collection: Project Documentation
- ❖ [Subcommunities of Michael Joyce Project:]
    - ➢ Series I. Works (this subcommunity consists of 39 collections, each containing the usually multiple files pertaining to one work)
    - ➢ Series II. Academic Career (this subcommunity contains the four subcommunities below)

[2] Items with an asterisk are gathered at the preliminary inventory stage. See http://www.hrc.utexas.edu/research/fa/

[3] The top level of this structure can be seen as displayed by DSpace at https://pacer.ischool.utexas.edu/handle/2081/289. The unfolding of the structure to "lower" levels can be followed by clicking on links. It should be noted that this structure is not fully populated, since additional in-hand material remains to be ingested and accretions to the Michael Joyce collection are anticipated.

    - Subseries A. Academic Works (this subcommunity contains 22 collections, each containing files pertaining to one “work”)
    - Subseries B. Administrative Material (this subcommunity contains 5 collections, each pertaining to one place of employment or grant project)
    - Subseries C, Conferences (this subcommunity contains 11 collections, each pertaining to a single conference attended by Joyce)
    - Subseries D. Teaching Material (this subcommunity contains 2 collections pertaining to Joyce’s teaching)
- Series III. Correspondence [currently unpopulated]
- Series IV. Storyspace (this subcommunity contains 2 collections pertaining to the StorySpace software written to support hypertext novel creation and rendering, one of which is currently unpopulated; this category has no paper collection counterpart)
- Series V. Journals and Appointment Books (this subcommunity currently contains 1 collection)
- Series VI. Personal (this subcommunity currently contains 1 unpopulated collection)
- Series VII. Works by Other Authors (this subcommunity contains the two collections below)
    - Subseries A. Published Works (this collection contains 8 works by other authors)
    - Subseries B. Works by students (this collection contains 6 works by students)

The detailed reticulation of this structure is characteristic of HRC descriptive practice and, in a way that is familiar to most archivists, mirrors (with one exception) categories that have come to be local terms of art within the repository; subseries usually appear in the folder list and do not normally receive narrative description in the conventional finding aids. What is interesting about this practice as articulated in the DSpace environment is the variability of the mapping of series and subseries onto the subcommunity and collection DSpace constructs, because in several cases a subseries or even a series may contain very few items and could clearly be mapped conveniently as a collection. But the descriptive practice calls for the separation of functional categories, even if each has only one member, and reflects finding-aid layout preferences in the reluctance to allow constructs that are children of a single series to be instantiated as different kinds of objects. Referring back to the explanation of the DSpace objects in question the practical reason for this can be suggested. It is necessary to create elaborate divisions and subdivisions in community mode to establish the overall structure, and then to generate collections when maximal homogeneity has been defined. This procedure also reflects in the high level of access required to create these hierarchically higher constructs an established practice where policies for arrangement conventions are in place and are enforced by practitioners with enhanced authority.[4]

For another example from a second large project for HRC in DSpace it is worth discussing the Arnold Wesker collection structure.[5] This structure was far more determined by an existing modern finding aid and an ample paper collection than was the Michael Joyce collection, but the consistency of HRC descriptive practice is clear in examining both digital collections, and the treatment of the Joyce collection in DSpace is further elucidated by comparing it with the categories in the online finding aids for the 1925-2000 segment of the Wesker collection and the 1958-2001 accretion.[6] An interesting quotation from the first Wesker finding aid gestures at the overriding repository-wide arrangement practice at the HRC: “In the process of boxing his papers for shipment to the Ransom Center, Wesker compiled a detailed listing of the contents, which is available for consultation. It provides a more fulsome description of the collection, complete with anecdotes and footnotes, and forms the basis for folder descriptions throughout. *However, the materials are not listed in the same order they appear in this finding aid*” (emphasis added). Because the digital Wesker collection was so large (5757 items ingested over three months’ total work by a team of three students, rough estimate 4 minutes’ processing per item) and the many files were ingested through the DSpace batch ingest process, no description was provided for individual items in the collection (collections were set up in advance with collection name and a brief description) apart from the file names serving as item titles. On the other hand, the overall collection was supplied with a quite detailed Biographical Sketch drawing on standard sources and a Scope and Content note specific to the digital part of the Wesker collection.

4.1.3 Representation as emergent digital archival practice: School of Information materials

Our experience with mapping archival descriptive practices onto DSpace has not been limited to the work with HRC’s established practice, as we have also established a number of collections that preserve digital archival materials of interest to the School of Information as documenting the history and technological evolution of the work of the School. In 2005 we began processing and ingesting collections that have so far included faculty publications and learning objects, digital videos documenting School events, computing lab tutorials created to help students with software and applications, and the School’s website. We had little to go on save a University-wide general administrative records schedule, the practices of the university archives at the Center for American History (now without an official university archivist and in any case in the early stages of developing practices for digital materials), and our own ability to ascertain the needs and requirements of our user community.[12]

Several of these projects have introduced specific descriptive practices that take advantage of DSpace features but that may not always be based on standard archival practice. Two examples come from the work that was done in 2005 in depositing the legacy works of four faculty members in the repository. Because

[4] A previous draft of the arrangement in a documentation file by Stollar Peters shows initial proposed structures that were not approved: https://pacer.ischool.utexas.edu/bitstream/2081/859/1/Arrangement.doc

[5] See https://pacer.ischool.utexas.edu/handle/2081/2220

[6] These are shown at http://www.hrc.utexas.edu/research/fa/wesker.html) and http://www.hrc.utexas.edu/research/fa/wesker.additional.html .

three of the four faculty members wished to deposit published papers, project archivists used the SHERPA/RoMEO database of publisher restrictions to determine what kinds of access could be provided and to make any restrictions explicit using the DSpace item.citation metadata element, which allowed the statement of any specific citation that restrictive publishers demanded.[7] In the case of the Andrew Dillon collection, which consisted entirely of published or unpublished papers, an effort was made in 2005 to experiment with extracting the subject terms for which DSpace provides metadata elements from the papers themselves using a tool from the TAPoR text analysis portal.[10] Subsequent projects, particularly those involved with materials lacking a useful controlled vocabulary, have made use of this process. Since DSpace now provides for the use of a controlled vocabulary, this process can potentially be incorporated into the workflow for setting up a corpus for ingest.

Further challenges from multimedia objects have elicited novel arrangement and descriptive practices for the collection of School of Information IT laboratory tutorials, a group of tutorials created by student laboratory assistants to support student use of software.[8] Discussions with IT staff responsible for designing the tutorial-creation process and with our Associate Dean determined that the tutorials were considered to be of both historical and technical value (we had our own motivations: since the tutorials demonstrate how to use many different software products, as digital archivists we must appreciate the long-term value of tutorials like this for understanding the objects created using these software products). Tutorials have a certain shelflife, and when the IT lab is no longer recommending or licensing a piece of software for student use, the tutorial is withdrawn. Because there is a large number of tutorials, it was decided to tackle the archiving problem in increments; so far one student working individually and two student teams in the digital preservation class have tackled the project. Almost all of the tutorials have a website-based structure and are accompanied by a video covering the same material in a different way. In the first year of work, the general parameters of the problem were defined and the collection set up; then archiving of the website materials for several tutorials was carried out, accompanied by a risk analysis for the relevant file formats, a history of tutorial creation at the School of Information, and documentary materials pertaining to the tutorial-creation workflow. In the second year of work, tutorial videos were archived, demanding experimentation with the creation of display formats to be archived as use copies along with the original raw formats and adoption of a convention for storing both.

Documentation materials, most of them couched as narrative reports, were archived in a set of collections under the subcommunity School of Information Tutorials Documentation, where there are now documentation collections from 2005, 2006, and 2007. In the course of our work with DSpace, the creation of these documentation collections—much like archival accession records in that they gather together miscellaneous information pertaining to the primary content collections, the processing to which they have been subjected, and any reports generated through working with them—has become a routine part of our practice for each fonds, as it allows the capture of any metadata that DSpace does not yet support and the explanation of pre-ingest processing done outside DSpace. It therefore provides to future digital archivists an account of specific technological steps that were taken and safely stores for future use information that can best be captured during initial processing. And because these collections are also collections in DSpace, they are searchable in the same way as are other collections. The documentation collections, then, represent a kind of meta approach to providing metadata to support the content of the repository, but at present they cannot be brought together with the materials to which they refer directly within DSpace except through the hierarchical grouping with the fonds to which they refer.

5. OPENING UP THE FINDING AID: DSPACE AND THE WORLD

Thus DSpace is not so constrained by its OAIS heritage that multiple and different descriptive practices are not possible, but all these practices are constrained to be presented to the world as DSpace is set up to do. So next we turn to the user's view of DSpace information structures. The DSpace representation of digital objects, by presenting the world with the opportunity to access them directly in several ways, has the effect of opening up both traditional and innovating archival representation in unanticipated ways. First it must be said that in the case of the HRC projects, unless the user is a physical visitor to the HRC, the actual digital objects cannot be accessed. This is due to a combination of reasons, most stemming from copyright concerns but others due to the archival terror of empty research rooms. Nevertheless, any online visitor to the Joyce or Wesker materials as represented in the School of Information's DSpace server will be able to access most of the metadata created by administrators at the subcommunity (biographical sketch, series descriptions) and collection (scope and content note) levels and by submitters at the item level (QDC displayed as short and long catalog entries).

But here we encounter the unanticipated: the fact that because these metadata are in DSpace and because DSpace was designed to incorporate some of the popular search capabilities of modern OPACS, the user can initiate a general free-text search across the whole repository, and will be presented with community, collection, and item hits. An Advanced Search even offers three-term Boolean search on keyword, author, title, subject, abstract, series, sponsor, identifier, and language QDC metadata elements. The user can also browse specific item-level fields (title, author, subject, "by date") both within a specific level or across any segment of the specific DSpace instance. Also at the community or collection level a free-text search can be used. Yet these searches unfortunately do not include descriptive material recorded at the community or collection level, but only the contents of a subset of the QDC metadata fields. It should be observed, however, that some of these searchable fields (notably abstract and description) can be artisanally populated with rich description and can yield interesting results.

This is rather more access than one would normally achieve using printed finding aids, and (as most of us have learned by now)

[7] For the SHERPA/RoMEO database, see http://www.sherpa.ac.uk/romeo.php.

[8] The School of Information's portal to current live tutorials is here: http://www.ischool.utexas.edu/technology/tutorials/. The archived materials for the tutorials project may be found at https://pacer.ischool.utexas.edu/handle/2081/334.

even EAD-encoded finding aids will not give this kind of access unless the encoding has been exploited for something more than online display.[4] DSpace has taken another step, evolving to open itself out to the utilities of Web 2.0, that breaks open the restricted view of the archival finding aid even further. First, through a partnership with the Open Archives Initiative, DSpace is configured to expose its QDC metadata to harvesting via OAI-PMH, which allows the aggregation by anyone with an OAI-PMH harvester of information about collections in OAI-compliant repositories. Second, through a partnership with Google, that search engine can, entirely respectful of any collection-specific DSpace access restrictions, search not only the QDC that OAI harvesting and the DSpace search functions provide, but also the metadata (here, biographical sketches, series descriptions, and scope and content notes) that are actually displayed on DSpace community and collection pages, thus providing in the Joyce and Wesker collection cases better access to the developed biographical sketch, series descriptions, and collection scope and content notes than DSpace can do itself.

What this means is that the careful design of online repositories for digital objects can not only allow both the generation and extraction of archival metadata at several hierarchical levels, but can support multiple kinds of searches on that metadata within the context of the repository and, through web services, expose that metadata to as yet unimagined uses. That metadata can include the documentation of archival process that has rarely been made available to researchers, and searches can additionally include the digital objects themselves where they have been made accessible. If and when archival descriptions and arrangements are superseded, the old versions can be retained as documentation and remain searchable as historic views of the materials. When the metadata is the system and the system is the metadata, archival description can eventually become archival on its own account.

6. ACKNOWLEDGMENTS

I thank all the students who have worked on digital archiving with me, whose work has helped to clarify these issues in the projects mentioned, notably Melanie Cofield, Lorraine Dong, Megan Durden, Leanda Gahegan, Kris Kelly, Thomas Kiehne, Sarah Kim, Dana Lamparello, Julia McGarey, April Norris, Gabriela Redwine, Lisa Schmidt, Nicole Soriano, Vivian Spoliansky, and Catherine Stollar-Peters. I also thank the University of Texas School of Information's Information Technology staff, notably Shane Williams and Sam Burns, who have provided DSpace support, and the many contributors to the open source DSpace project all over the world.

7. REFERENCES

[1] Cohen, Daniel and Rosenzweig, Roy. Digital History (Philadelphia, University of Pennsylvania Press, 2006), 228.

[2] Describing Archives: A Content Standard (Chicago, Society of American Archivists, 2004).

[3] Duff, Wendy. Will Metadata Replace Archival Decription: A Commentary. Archivaria 39 (Spring 1995): 33-38.

[4] Gilliland-Swetland, Anne. Popularizing the Finding Aid: Exploiting EAD to Enhance Online Discovery and Retrieval in Archival Information Systems by Diverse User Groups. In Daniel Pitti and Wendy Duff (eds.), Encoded Archival Description on the Internet (New York, Haworth, 2001): 199-225.

[5] Hedstrom, Margaret. Descriptive Practices for Electronic Records: Deciding what is Essential and Imagining What is Possible. Archivaria 36 (Autumn 1993): 53-63.

[6] Jones, Richard, Andrew, Theo, and McColl, John. The Institutional Repository (Oxford, Chandos, 2006), Chapter 4.

[7] Kinner, Jason A. DSpace History System. DOI=http://simile.mit.edu/reports/dspace-history/design.pdf

[8] Lazorchak, William M. The Ghost in the Machine: Traditional Archival practice in the Design of Digital Repositories for Long-Term Preservation. Master's Thesis. School of Information and Library Science, University of North Carolina-Chapel Hill, 2004.

[9] Lucas, Lydia. Efficient Finding Aids: Developing a System for Control of Archives and Manuscripts. American Archivist 44 (1981): 21-26.Yakel, Elizabeth. Archival Representation. Archival Science 3 (2003): 1-25.

[10] Norris, April. 'Putting More Stuff in the Thing': The Completion of the Andrew Dillon DSpace Digital Collection, pp. 10-11. DOI=http://hdl.handle.net/2081/1179

[11] Roe, Kathleen D. Arranging & Describing Archives & Manuscripts (Chicago, Society of American Archivists, 2005), 46 (Figure 4-1).

[12] Sevcik, Edward A. Shearing Layers of Trust: Considerations for Digital Authenticity at a Research Institution that also serves as a University Archives. Master's Thesis. School of Information, University of Texas, 2007.

[13] Stollar Peters, Catherine. When Not All Papers are Paper: A Case Study in Digital Archivy. Provenance 24 (2006): 22-34.

[14] Yakel, Elizabeth. Archival Representation. Archival Science 3 (2003):1-25.