Archival IR: Applying and Adapting Information Retrieval Approaches in Archives and Recordkeeping Research

(DRAFT—please do not cite or redistribute)

Jonathan Furner and Anne J. Gilliland

Abstract: In this chapter, the prospects for archival information retrieval (IR) as a research area within the archives and recordkeeping domain are reviewed with the aim of encouraging its application. IR is characterized as a body of techniques with wide applicability, but with relatively little influence, historically, on the design of systems offering intellectual access to archives and records. Significant terminological differences (and overlaps) are noted between the IR field and the data archiving, and archives and recordkeeping domains. The principal concepts and objectives of IR are summarized, and the trajectory of archival IR outlined, with a focus on myths, challenges, and recent developments. XML retrieval is identified as a primary locus for researchers in archival studies to participate in the design and development of the next generation of IR systems. It is suggested that potential advances in archival IR—such as helping users to find previously unknown and possibly "smoking gun"–type documents; establishing the meaningful absence (as opposed to the presence) of documents; and exploiting multiple types and sources of metadata—may find wider application in other domains such as litigation support systems, news retrieval, audiovisual archives, data mining, and digital asset management.

Introduction

Information retrieval (IR) is the name that has been used since the 1950s to refer to an interdisciplinary field of inquiry that draws its methods from computer science, library and information science, linguistics, statistics, and psychology.[i] Researchers in IR seek to improve our understanding of the ways in which people can find, among large quantities of resources that contain information (broadly defined), resources of the particular kinds that they want. Over the course of the past seven decades, the scope of the field has been extensively and variously delineated by researchers with wide-ranging interests, but definitions have frequently included, as objects of study, the beliefs, goals, values, intentions, actions, and products of the following groups:

- *information seekers* (a.k.a. *searchers*), i.e., those who are looking for information;
- *IR systems designers*, i.e., those who devise and build systems and services (manual or automated, analog or digital, stand-alone or networked) that provide assistance to information seekers; and
- intermediaries such as *indexers*, *catalogers*, and *processors*, who pre-process information resources in such ways as to make them more accessible to seekers—for example, by identifying terms, headings, codes, or descriptors of some kind, to use as descriptive *labels* for resources (or for classes of resources), and by creating more- or less-complete *representations* of, or surrogates for, resources.

Often IR is conceived rather more narrowly as the art and science of producing and improving upon computerized retrieval systems (a.k.a. *search engines*) that help information seekers both to find more of the information that is wanted, and to avoid more of the information that is not.[ii] Some IR research is dedicated to the creation and/or implementation of such systems, or of particular system components such as user interfaces;[iii] other research involves the measurement and evaluation of the performance (a.k.a. *retrieval effectiveness*) of such systems;[iv] and yet other research constructs theory that seeks to explain why systems of one kind perform to a higher standard than those of another.[v]

IR techniques have been widely applied in diverse settings. At the time of writing, the world's most widely used IR system is the Web search engine Google. The Microsoft Windows and Apple Mac operating systems both incorporate search engines that allow computer users to find relevant resources in their own personal collections. In libraries both physical and digital, patrons use OPACs (online public access catalogs) to identify desired materials. Each of these kinds of automated IR system has a long and more-or-less illustrious history.

In contrast, IR techniques have not been so widely applied in the provision of access to records and archives. Prior to the 1980s and the widespread implementation of electronic recordkeeping, archivists in countries and sectors with strong registry traditions relied upon a centralized registry office or system that structured workflow and identified, classified, controlled, and sometimes eliminated records generated by bureaucratic activity prior to those records being received by the institutional archives. These registry systems thus provided the fundamental infrastructure for manual information retrieval for both active and archival records. In the absence of registry systems—for example in the United States where they were never widely adopted, or where archival resources resulted from personal activity—archivists relied upon their own knowledge of archival holdings, and of their associated filing schemes and finding aids, in order to meet users' expressed needs. As automated recordkeeping was increasingly implemented and hierarchical information flows and centralization of recordkeeping activities were replaced by network structures, registry systems increasingly broke down, as did many other forms of systematized bureaucratic filing systems. A concomitant need arose for records creators and archivists to implement robust IR mechanisms for the increasingly voluminous products of institutional recordkeeping. Commercial developers addressed this need by designing electronic records management (ERM) systems, electronic document management (EDM) systems, digital asset management (DAM) systems, and other forms of resource management systems for use within and across bureaucratic settings. Developments of this kind were accompanied by bursts of enthusiasm in the 1980s and 1990s for "archival informatics" in general, and for subject indexing of archival holdings in particular.[vi] However, the notion of or need for "archival IR"—i.e., the adoption and adaptation of IR concepts and techniques to address specific archival and recordkeeping needs and problems—remained substantively un-addressed within archival studies.[vii]

Historically, when those in the archives and recordkeeping domain *did* talk about archival IR, one or more myths, or (at best) only partial truths—typically about what is "classic IR," and how it is not applicable in archival contexts—were frequently perpetuated. A list of these might include the following:

- IR is all about the provision of access to *information*—whereas archives are all about the preservation of the products (e.g., records) of bureaucratic and personal activity as *evidence* of that activity.
- IR is a post hoc set of activities, conducted *after* resources have been acquired by a library or other repository—whereas long-term access considerations for records need to begin from the moment a recordkeeping system is being designed by or for a records creator, and to continue *throughout the life* of those records.
- IR is primarily about helping information seekers gain item-level *access* to resources—whereas archival processing is primarily about describing, explaining, and presenting the products of active and archived recordkeeping in context, in order to facilitate primary and secondary *use, and re-use*. That context comprises the various agents, activities, mandates and functions associated with those products and the relationships between them in and through time.[viii]
- IR places high value on improvements in the quality of *item-level* subject indexing—whereas archivists focus on *collection-level* description to ensure that items are always retrieved in context, and eschew subject description and retrieval based on corporate, personal, or place names because of the high incidence of inconsistencies and historical and cultural variations in the choice and form of such names. At any rate, archivists could never cost-effectively become involved in detailed, item-level subject indexing of their holdings, simply because of the magnitude of the manual effort apparently required.
- IR is good only for searchers whose information needs can be expressed as *topical* subjects—whereas archival holdings are typically described *provenancially*, usually by personal and organizational names that are also subject to historical and cultural variation, and that are notoriously inconsistently applied by records creators, archival processors, and end-users.
- IR relies on resource descriptions made up of statements of certain observable characteristics of bibliographic materials, such as title, author's name, and publication date, and hence is primarily the domain of *libraries* of such materials—whereas records and other archival materials usually lack such bibliographic characteristics. Moreover, archivists are wary about what they would have to "give up," "shoehorn," or add on to their descriptive processes in order to be able to take advantage of classic IR techniques.
- IR is primarily about making advances in the design of algorithms to be followed by *machines*—whereas archival retrieval invariably relies on the unique talents, specialized knowledge, and prodigious memories of the *humans* who take care of archival holdings, especially those who have been closely engaged with processing and providing reference services to particular holdings.
- IR is good only for *digital* resources—whereas the majority of archival holdings are not yet in digital form, and some holdings may never be.

- IR is good only for *textual* resources—whereas archival holdings often include non-textual materials such as photographs and recordings.

Clearly, many of these claims about the lack of applicability of IR to archival settings are based upon outdated notions about contemporary IR techniques, and outmoded conceptualizations of archives and their descriptive practices today. Several are demonstrably being debunked not only by advances in contemporary recordkeeping systems, but also by broad-based archival developments. These latter include the following: mass digitization of holdings; creation of metadata for each digitized item; generation of searchable full-text versions of digitized textual documents; creation of linked data; development of standards for the structure of authority files; construction and sharing of standardized authority files; and increasing reliance on full-text search engines to provide enhanced searching of online finding aids generated by "slimline" processing procedures such as More Product, Less Process (MPLP).[ix] The end result—networks of large-scale online archives, implemented at intra- and inter-institutional levels and in centralized and federated forms, that are making available born-digital as well as digitized content, together with collection- and item-level metadata—present a compelling case for developing a robust agenda for archival IR research that will support and enhance use of these archives.

In this chapter, through an exposition of classic IR ideas and approaches and contemplation of the conditions and needs of twenty-first century archives and recordkeeping, we argue that IR provides a key set of concepts and methods for those who seek to enhance archival access and use. Furthermore, because of its conceptual and temporal complexities, the archives and recordkeeping domain offers IR researchers opportunities to probe some of these complexities further, and provides a rich vein for nuanced development of the IR field as a whole. Our aim with this chapter, then, is to review both the actuality of, and the potential for, the application of IR approaches in archival studies research—revisiting the above myths in the process, as appropriate. After a brief note about how potentially confusing terminological overlaps and differences between and within the IR field and the archives and recordkeeping domain might be addressed, we present a broad outline of the conceptual framework that we shall be using to situate archival IR, simultaneously in the field of IR as classically understood and in archival studies. We then review a selection of prior and recent approaches to archival IR, and speculate about the future prospects for archival IR, before drawing some final conclusions.

A Note on Terminology

That there has been almost no historical interaction between the archival and IR fields[x] is immediately evident from the terminology used in each context. Some terms are used in both fields, but denote different concepts in each; some concepts are shared by both fields, but are denoted by different terms in each. Such conflict importantly reflects deeper conceptual differences between the roles, procedures, and points of engagement of those who have historically developed the IR field and those in the archives and recordkeeping domain (and, more recently, in the related field of data archiving). Unlike

other fields associated with the "information sciences" whose scholars characteristically have looked to the early 20th-century *documentation* movement for inspiration,[xi] archival studies is concerned specifically with aligning *records*, their users, and their uses from the moment they are imagined in the design of a records (a.k.a. recordkeeping) system, and for as long as the resources generated by those systems continue to exist, whether in their original setting or after transfer to a physical or digital archives. As a result, IR considerations must begin at the point of the creation of the original records system, and must continuously be attuned to and aligned with the needs, behaviors, and practices of the various kinds of users who wish to access the records system and its content over time. In this respect, archives and recordkeeping as a domain has close ties to the emerging field of data curation, as well as to institutionally-based professions, such as librarianship and museum collections management, that still largely rely upon post hoc information processing to support information retrieval.

Much of the canonical IR terminology can be traced at least to the Cranfield tests—a series of influential experiments, conducted by the British librarian Cyril Cleverdon in the 1960s, in which the impact on retrieval effectiveness of several different methods of indexing was measured in a controlled setting[xii]—and, further back, to Calvin Mooers' first use of the term "information retrieval" in 1950.[xiii] Meanwhile, archives and recordkeeping terminology has evolved according to the field's own historical and cultural trajectories over centuries, and arguably with rather less agreement than can be found in IR. However, certain archival terminology is now too embedded in national and international standards for archivists to contemplate change. For archival IR to gain traction, the meanings of terms must be clarified, distinguished, and mapped, so that confusion (both internal and external) may be avoided. This is certainly not a problem that is unique to this context, but rather is illustrative of the processes that have to occur when any method is adopted, adapted, and internalized within a new domain, and especially if it is hoped that outcomes will be fed back into the parent field.

The terminological difficulties that plague any discussion at the intersection of archival studies and information studies may be summarized as follows:

1. The term "information," notoriously, is used in a large number of different ways, to the extent that there is seldom much agreement on a preferred sense even within relatively small user groups, let alone across entire disciplines, professions, or national traditions. (a) Some find it possible to distinguish *objective* senses of information-as-signifier (e.g., marks on a page) from *subjective* senses of information-as-signified (e.g., meanings ascribed to marks). (b) At the same time, some commonly see a benefit in distinguishing between information that has *ultimate* value (e.g., the content of a document that precisely meets an information seeker's needs) and information that has merely *instrumental* value (e.g., the metadata that comprise a description of the document sought). (c) Third, some distinguish between the information supplied by a document in virtue of its *content* (e.g., information about subject matter) and that supplied in virtue of its *existence*, *form*, and/or *structure* (e.g., information about provenance or context). Indeed, some of those in the last camp will assess the informational value and the evidentiary value of a

document separately, implying that information and evidence are different things—or, at least, that information-as-evidence is a discrete species of the information genus.

In the field of information retrieval, it is typically assumed that, even if the information sought is essentially subjective, it is valid to base retrieval on inferences drawn from observation and measurement of objective information; and that metadata are of clear utility in serving as surrogates for documents in the collections searched. The information/evidence distinction, however, is not one that routinely impinges on IR systems design. In archival studies, on the other hand, it is this third distinction that historically has been treated as by far the most important, to the extent that external debates about the nature of information have remained tangential to the interests of archivists and archival theorists alike. All else being equal, the evidentiariness of a given archival record is likely to be valued more highly than its informativeness, and it is the nature of evidence (not information) that usually exercises philosophically-inclined minds in the archival field.

2. Another term in ordinary usage has several remarkably different technical senses, not just in archival studies and IR, but also in the related field of library science. The English term "records" has been in common use since the fourteenth century in referring to documents of a particular kind—viz., those that serve as archival evidence.[xiv] Only since the late 1950s has the term "record" been used also to mean a unit of information (superseding the slightly earlier use of "item" for the same purpose). The use of "record" in place of "entry," e.g., in the phrase "catalog entry," is more recent still, dating from the mid-1960s (whereas its precursor can be traced back at least to the sixteenth century). The present situation, then, is one of no little confusion, in which a single term does triple duty as the name for, (a) in archives and records management, and in recordkeeping more broadly, a class of documents,[xv] (b) in computer science, a class of descriptions (of objects in general),[xvi] and (c) in librarianship, a different class of descriptions (of documents in particular).[xvii]

3. "Information" and "record" are not the only sites of contesting claims on semantic resources. For example: In IR, "collection" is normally used as it is in library science, to refer generally to any "gathering of documents assembled on the basis of some common characteristic,"[xviii] regardless of whether that shared characteristic is provenance. In archival and records terminology, on the other hand, "collection" is sometimes used to refer specifically to any thematically-based or other purposive gathering of documents assembled *without* regard to provenance (e.g., by collectors or by collecting archives),[xix] other times to refer specifically to any gathering of documents *with* shared provenance (e.g., by institutional archives),[xx] and yet other times (often in the plural form "collections") to denote the holdings of a given repository (e.g., archives, historical collections, and library special collections).[xxi]

4. Similarly little consensus, cross- or intra-disciplinary, exists on the appropriate generic term to use in referring to the kind of things that may potentially form the contents of archival repositories—i.e., what we (until this point in the present paper)

have been calling "documents." In the most recent glossary published under the auspices of the Society of American Archivists, Richard Pearce-Moses settles on "material" as "an encompassing, generic term to describe the broad variety of items that an archives might collect, regardless of medium, format, or type," noting that this is done in order to "avoid connotations carried by terms such as record, document, or object," and that in this sense "'material' is roughly synonymous with 'resource.'"[xxii] This compromise, however, tends to reflect the somewhat anomalous situation in the United States—where, unlike in many other regions of the world, the archival domain encompasses both institutional archives and historical manuscript and library special collections.

At the same time, the candidacy of "document" (let alone the briefly popular "document-like object") as the appropriately generic term is impugned by tendencies in some quarters to interpret it as denotative only of textual, non-official, or even non-evidentiary things, or only of specific instances of records, such as medieval charters. In recent years, the term "resource" has emerged from the digital library community as a contender for naming this top-level category of things—including both records and (some) non-records—that may be collected, described, sought, and discovered.[xxiii] In the digital-library domain, however, a sharp distinction is often drawn between (on the one hand) resources and (on the other) metadata. It should be noted that the archival community makes such a distinction only between records (or materials) and *descriptive* metadata. In other words, archivists are among those who are careful to acknowledge that, depending upon the context, metadata of non-descriptive types may themselves also be considered as records.

In contemplating various possible strategies for tackling the terminological problem for the purposes of writing this chapter, we felt that it was important to resist the temptation of presenting it as one of a simple dichotomy between IR on the one hand, and archives and recordkeeping on the other, since even within-field consensus about the meanings of terms is not complete. We considered three alternative strategies, as follows:

1. One strategy would be to undertake a mapping between the canonical terminology used in IR and that used in archives and recordkeeping, but this approach runs the risk of inadequately representing the nuances and historical shifts that have taken place within each context. In the archives and recordkeeping domain, for example, where one is dealing with differing professional formations in different jurisdictions, arriving at terminological consensus has been notoriously difficult. There have been extensive debates about the definitions of and relationships between such fundamental terms as "record" and "archive," and even about the scope of the term "records management," in the technical committees that oversee the development and revision of ISO records management standards.[xxiv] Similarly, the International Council on Archives (ICA), which promulgates the ISAD suite of standards for archival description,[xxv] has been unable since 1988 to bring a dictionary or glossary to publication.[xxvi] We pondered whether it might be possible to identify multiple discrete positions or perspectives in each area (e.g., in IR, traditional and progressive, objectivist and subjectivist; and in archives and recordkeeping, life cycle, records

continuum, and digital curation) that are each characterizable by more-or-less stable definitions of how each term is being used within that perspective. Such an approach would certainly make for interesting research in its own right, but we felt that it was too large and complex an endeavor for what we were attempting to achieve with this chapter. A variant approach might have been to construct, for each term, a list of the properties that a given entity must have if it is to be denoted by that term in a given domain, but again, given the distinct differences in perspectives identified above, that was also deemed to be too complex an approach for this chapter.

2. A second strategy would be to use purposively disambiguated and non-aligned language to present our exposition. We attempted this strategy in our initial drafts of this chapter, but felt that it ended up diminishing the canonical aspects of traditional IR, and engaged us in all sorts of terminological contortions that only added confusion into an already complex discussion.

3. A third strategy, and the one that we ultimately pursued, would fall in the space between the previous two: retaining canonical IR terminology as refracted through the lenses of the *DCMI Glossary*,[xxvii] *ISO 5127:2001 Information and Documentation — Vocabulary* and *ISO 25964-1:2011 Information and Documentation — Thesauri and Interoperability with Other Vocabularies — Part 1: Thesauri for Information Retrieval* but also employing terms that are central to archival studies, including some of the terms defined in two standards that are now being widely adopted in digital archives and recordkeeping, and in data curation. *ISO 30300:2011 Information and Documentation — Management Systems for Records — Fundamentals and Vocabulary*, seeks to update and reconcile terminology used in various prior ISO standards for records management (RM).[xxviii] While nominally labeled RM, it has been strongly influenced by records-continuum conceptualizations of recordkeeping that encompass archival activities, and currently represents the most expansive (albeit incomplete) consensus of different records management and archival constituencies. *ISO 14721:2012 Space Data and Information Transfer Systems — Open Archival Information System (OAIS) — Reference Model*, developed by the Consultative Committee for Space Data Systems (CCSDS) and initially adopted in 2003, is directed toward data archiving and is intended for use across diverse domains. It is being implemented by many digital archives, preservation, and curation initiatives as the underlying framework supporting the ingest, management, and retrieval of a diversity of digital content, and provides something of a bridge between the archives and recordkeeping domain and broader constituencies concerned with retrieval of "archived" digital information objects.[xxix]

Table 1 lists some of the key terms used in the five standards noted above. Beyond providing the reader with the meanings of terms in their different contexts, the table clearly illustrates the different preoccupations and perspectives of the areas that need to be aligned, or at least understood, if archival IR is to become more widely pursued.

Table 1. Definitions of selected terms in five standard glossaries.

Note: Term/definition pairs marked with an asterisk (*) are those used in this chapter.

1. Resources

1.1 document

DCMI Glossary	**resource**: anything that has identity …
	***information resource**: any entity, electronic or otherwise, capable of conveying or supporting intelligence or knowledge; e.g., a book, a letter, a picture, a sculpture, a database, a person
	document-like object: any discrete information resource that is characterized by being fixed (i.e., having identical content for each user); … includes text, images, movies, and performances
ISO 25964-1:2011	**document**: any resource that can be classified or indexed in order that the data or information in it can be retrieved
ISO 5127:2001	**document**: recorded information or material object which can be treated as a unit in a documentation process
	unit of description: document and its parts or aggregations treated as an entity
ISO 30300:2011	***document**: recorded information or object which can be treated as a unit
ISO 14721:2012	-

1.2 records

DCMI Glossary	-
ISO 25964-1:2011	-
ISO 5127:2001	**record**[2]: document created or received and maintained by an agency, organization, or individual, in pursuance of legal obligations or in the transaction of business
ISO 30300:2011	***record(s)**: information created, received, and maintained as evidence and as an asset by an organization or person, in pursuit of legal obligations or in the transaction of business
ISO 14721:2012	**content information**: a set of information that is the original target of preservation or that includes part or all of that information. …
	archival information package (AIP): an information package, consisting of the content information and the associated preservation description information (PDI), that is preserved within an OAIS

1.3 archives

DCMI Glossary	-
ISO 25964-1:2011	-
ISO 5127:2001	**archives**[1]: records[2] of the same provenance accumulated by an organization or person in the course of the conduct of affairs, and preserved because of their enduring value
ISO 30300:2011	***archives**[1]: records maintained for continuing use
ISO 14721:2012	-[xxx]

2. Collections of resources

DCMI Glossary	-
ISO 25964-1:2011	-
ISO 5127:2001	***collection**[2]: gathering of documents assembled on the basis of some common characteristic

ISO 30300:2011	-
ISO 14721:2012	**archival information collection** (AIC): an archival information package whose content information is an aggregation of archival information packages

3. Metadata about resources

DCMI Glossary	**metadata**: in general, data about data; functionally, structured data about data; … includes data associated with either an information system or an information object for purposes of description, administration, legal requirements, technical functionality, use and usage, and preservation …
	record: some structured metadata about a resource, comprising one or more properties and their associated values
	metadata record: a syntactically correct representation of the descriptive information (metadata) for an information resource …
ISO 25964-1:2011	**metadata**: data that identify attributes of a document, typically used to support functions such as location, discovery, documentation, evaluation, and/or selection
ISO 5127:2001	**record**[1]: set of data on one person or object, selected and presented for a predefined specific purpose
	description[1]: … results … [of operations] including capturing, analyzing, organizing and recording of data on documents in order to ensure their identification and control
ISO 30300:2011	***metadata**: data describing context, content, and structure of records and their management through time
ISO 14721:2012	**metadata**: data about other data
	preservation description information (PDI): the information which is necessary for adequate preservation of the content information …

4. Users of resources

DCMI Glossary	-
ISO 25964-1:2011	-
ISO 5127:2001	***information user**: utilizer of infrastructures, services, or material offered by information centers
ISO 30300:2011	-
ISO 14721:2012	**consumer**: the role played by those persons, or client systems, who interact with OAIS services to find preserved information of interest and to access that information in detail
	designated community: an identified group of potential consumers who should be able to understand a particular set of information …

5. Resource-description processes

DCMI Glossary	***indexing**: the process of evaluating information entities and creating terms that aid in finding and accessing the entity …
ISO 25964-1:2011	**indexing**: intellectual analysis of the subject matter of a document to identify the concepts represented in and allocation of the corresponding index terms to allow the information to be retrieved
ISO 5127:2001	***description**[1]: operations … including capturing, analyzing, organizing and recording of data on documents in order to ensure their identification and control
	indexing: denotation of the content or form of a document by means of words[1], phrases, or notations[2], according to the rules of an indexing language
ISO 30300:2011	**indexing**: establishing access points to facilitate retrieval
ISO 14721:2012	-

6. Resource-discovery processes

DCMI Glossary	***resource discovery**: the process through which one searches and retrieves an information resource
ISO 25964-1:2011	***information retrieval**: all the techniques and processes used to identify documents relevant to an information need, from a collection or network of information resources
ISO 5127:2001	**information retrieval**: process of recovering specific information[1] or information[2] from a store
	document retrieval: process of recovering specific documents from a store
ISO 30300:2011	-
ISO 14721:2012	-

7. Resource-management organizations and systems

DCMI Glossary	-
ISO 25964-1:2011	-
ISO 5127:2001	**archives**[2]: organization or part of an organization responsible for selection, acquisition, preservation, and availability of one or more archives[1]
ISO 30300:2011	***archives**[3]: an organization, agency, or programme responsible for selecting, acquiring, preserving, and making available archives[1]
	***records system**: information system which captures, manages, and provides access to records over time
ISO 14721:2012	**archive**: an organization that intends to preserve information for access and use by a designated community
	open archival information system (OAIS): an archive, consisting of an organization … of people and systems, that has accepted responsibility to preserve information and make it available for a designated community

8. Resource-discovery systems

8.1 retrieval system

DCMI Glossary	**discovery software**: a computer application designed to simplify, assist, and expedite the process of finding information resources
	search engine: a utility capable of returning references to relevant information resources in response to a query
ISO 25964-1:2011	-
ISO 5127:2001	***retrieval system**: system allowing access to representations of documents, their addresses in a collection[2], and the documents themselves
ISO 30300:2011	-
ISO 14721:2012	**access aid**: a software program or document that allows consumers to locate, analyze, order, or retrieve information from an OAIS

8.2 finding aid

DCMI Glossary	-
ISO 25964-1:2011	-
ISO 5127:2001	***finding aid**: retrieval system produced to establish control over records[2] …
ISO 30300:2011	-

ISO 14721:2012	**finding aid**: a type of access aid that allows a user to search for and identify archival information packages of interest

IR: A Conceptual Framework

1. Intellectual access and resource discovery

By implementing an IR system, giving people the opportunity to use it, and providing them with tools to understand the resources they are accessing, the stewards of a given collection of resources are said to provide *intellectual access* to the resources in that collection. Intellectual access may be distinguished from at least two other kinds of access to resources:

- the *physical access* that is given to those who are able to interact physically with the resources themselves (or, according to some conceptions, with the content of the resources); and
- the *legal access* that is given to those who are permitted, under the laws or regulations of the relevant jurisdiction or stipulations of a donor or institution, to make use of the resources in certain prescribed ways.

Intellectual access is access of the kind that is enjoyed by those who are successful in finding the resources that they want. IR systems help to provide such access to the extent that they assist the user in the process of *resource discovery*—i.e., in the process by which the searcher identifies resources that, to a greater or lesser degree, match whatever criteria the searcher uses, at any given time, to judge resources' desirability (or "relevance").

Conventionally, the primary means by which IR systems support resource discovery has been through the effective operation of, among others, two system components: one that generates more- or less-detailed descriptions of resources, and one that generates rankings of resources on the basis of the degree to which their descriptions match the criteria specified in users' *queries* (a.k.a. search statements; expressions of users' information needs). In other words, an IR system typically includes an *indexing* mechanism, which takes care of resource description, and a *relevance ranking* mechanism, which ranks resources in order of their descriptions' degree of similarity, or probability of relevance, to individual queries.

2. Resource description, metadata, and authority control

In library science, the process of resource description has been known historically either as cataloging or as indexing, in rough accordance with the nature of the resources being described: typically, book-length resources are cataloged, while resources such as journal articles, which may be conceived as parts of larger wholes, are indexed. From the 1990s onwards, the term *metadata* has come to be used more frequently to denote the content of the products of resource description; and catalogers and indexers are now often said to be in the business of assigning metadata to resources.[xxxi] In the archives and recordkeeping

domain, in contrast, the library terminology of cataloging and indexing is less frequently used: in recordkeeping, bureaucratic records would typically be *classified* and/or filed according to a pre-established scheme; and in archives, content would be collectively and hierarchically arranged and described in the course of archival *processing*, and the descriptive product would be a finding aid. Moreover, to the extent that the terminology of metadata has come to be used in archives and recordkeeping, the term refers generally to *all* data relating to an information resource, its creation, management and use that are generated over the course of its life, not just descriptive data intended specifically to facilitate discovery (a.k.a. descriptive metadata). Much of this metadata, even if it is not created expressly for descriptive purposes, can nevertheless be exploited in IR.

Metadata may be created or assigned either manually or automatically.[xxxii] Since the 1960s, the comparative quality of manual and automatic metadata creation in terms of its utility for IR has been debated frequently and at length, usually with a shared understanding of metadata "quality" that gives most weight to the utility with which assigned metadata allow searchers to discriminate between more-relevant and less-relevant resources.[xxxiii] Assigning metadata to represent the topical subjects of the contents of resources—known variously as subject description, subject cataloging, subject indexing, or sometimes simply indexing—is often considered independently as an especially problematic case. The influential Cranfield tests in the 1960s appeared to show that the products of simple methods of automatic subject indexing, relying only on the extraction of meaningful words from pre-existing titles and abstracts of resources, are at least as valuable as those of some more-complex (and therefore costlier) methods of automatic subject indexing, and as those of manual subject indexing.[xxxiv] Later studies appeared to demonstrate the relatively high quality of sets of index terms extracted automatically from the full-texts (instead of the titles and abstracts) of resources, and of sets of index terms obtained by carrying out certain kinds of statistical analyses of the frequency of occurrence of term-types, both in individual resources and in whole collections.[xxxv]

Notwithstanding the empirical evidence that current methods of automatic subject indexing are highly effective, libraries have persisted in dedicating non-trivial amounts of time and money to the manual assignment of index terms selected from predefined lists (i.e., *authority files*), such as the list of subject headings authorized by the Library of Congress (LCSH).[xxxvi] The justifications given for doing so typically invoke a rationalist argument to the effect that, when index terms are "controlled" in the way that LC subject headings are, (a) the chances are increased that a searcher will choose a search term which matches an index term assigned to a wanted resource, and/or (b) the searcher is able more easily (and ultimately more effectively) to browse among the classes of resources represented by individual index terms. With the adoption of national and international standards for describing archival content such as DACS, EAC-CPF, and ISAAR(CPF), archives are investing more than they have in the past in assigning subject headings within their finding aids.[xxxvii] However, wary of the costs involved and also the problems of negotiating the idiosyncratic, archaic, or technical language used in many archival resources, many archivists remain unconvinced that vocabulary control of *subject* terms is effective in supporting user access. Instead, they believe that keyword

searching of full-texts of finding aids and digital or digitized documents, when combined with searches of (a) *provenancial* access points (i.e., creator or collector), and (b) indicators of collection structure or arrangement, has a greater likelihood of producing a match between the terminology used by the searcher and that used in relevant resources.

Establishing control over vocabularies of index terms involves activities of several kinds:[xxxviii]

- identifying the *semantic relationships* that exist among terms: equivalence relationships, between terms that have similar meanings; hierarchical relationships, between terms that are broader and narrower in scope; and associative relationships, between terms whose meanings are related in some way;
- identifying a set of *entity-classes* (such as "Agent," "Action," "Object," "Concept," "Event," "Place" …), each of which is made up of a discrete set of paradigmatically related terms; and
- creating a set of *authority data* for each term, in which representations of semantic term–term relationships and (potentially) further metadata about the term are recorded.

Each of the resulting sets of authority records for terms in a given entity-class forms an authority file. Since provenance is traditionally the primary access point for archival resources, the recent development of the ICA standards for corporate, personal and family name authorities (ISAAR(CPF)) and for recordkeeping functions (ISDF) has led to new efforts to create and share archival and recordkeeping authority data.[xxxix] This is occurring most notably in Europe where historically dynamic national boundaries and the movement of unique records during and after conflicts and conquests has often resulted in records of the same provenance or relating to the same region being distributed across multiple repositories, frequently within different national jurisdictions. Records pertaining to the same region, population, or bureaucratic function may also be created using different languages and terminology, depending upon the ruling administration.

The primary benefits of maintaining authority files may be summarized as (a) the potential for sharing the files among distributed users, as is done on a global scale in the case of the Library of Congress's name and subject authorities, for example; (b) if the files are co-constructed by multiple collaborating institutions, a reduction in the average amount of authority work to be done in any single institution, and potentially an increase in the scope of the information included in the authority file because each contributor may be in possession of different information; (c) the elimination of redundancy in resource descriptions, with the data about a given term recorded only once instead of in every metadata statement in which it is included as an access point; and, in the archival context in particular, (d) support for the disambiguation of similar entries relating to different names or functions, and for the collocation of variant entries relating to the same name or function, thus assisting the user in identifying relevant archival resources.

More recently, system designers have explored the idea that metadata might usefully be supplied by "the crowd" of end-users, in addition to, or even rather than by professional

catalogers.[xl] Some systems allow for the direct "tagging" or annotation of resources by end-users; others log searchers' queries, treat their "click-throughs" as implicit relevance judgments, and assign as index terms the search terms that are most frequently used by those searchers who click through to view the resource in question. "Recommender" systems—which log the decisions made by end-users for example to cite, view, download, or purchase particular resources, and treat as metadata the resulting user profiles, and/or the features of resources that are closely related in co-occurrence networks—may similarly be conceived as implementing a variant of indirect crowdsourcing.[xli]

3. "Relevance" ranking

The "scare quotes" around the first word in the subheading, above, are intended to highlight the nature of the conceit at the heart of the IR process, which is that it is possible for IR systems accurately to determine the degree of relevance of any given resource to any given searcher at any given time. We should rather say that IR systems vary in the mean effectiveness with which they are capable of distinguishing resources that are more *likely* to be relevant from those that less likely to be—on the basis of statistical analyses of the frequency of occurrence of certain features (for example, terms or links) in individual resources, in whole collections, in assigned metadata, and in queries. In other words, at best, we can be optimistic that systems can estimate *probabilities* of relevance; divining the *actual* relevance of a resource to a searcher at time *t*—which is an entirely subjective matter—is (currently, at least) beyond the capability of any machine.

Nevertheless, once we have persuaded ourselves that even if this poses a problem for the theorist, it does not for the practitioner, two important ideas—(a) that resources may be ranked in order of the likelihood that they are relevant under prevailing conditions, and (b) that systems may be evaluated by determining how well, on average, they can predict the preference orderings of users—will be grasped straightforwardly enough. Two complementary measures of retrieval effectiveness are especially well known: *Recall* is the proportion of relevant records that are retrieved; *precision* is the proportion of retrieved records that are relevant. One fairly conventional way of summarizing the effectiveness of a given ranking mechanism is to plot the mean precision scores obtained at a series of incrementally increasing levels of recall (e.g., 0.1, 0.2, 0.3, …, 1.0).

4. Resources and representations

As we discussed in "A Note on Terminology" above, a distinction has often been drawn in the digital-library community, as well as in others that overlap similarly with the IR domain, between *resources*—i.e., the materials whose informational and/or evidentiary content is what is sought by searchers—and *metadata*, i.e., the descriptions or representations of resources with which searchers' queries are compared (*descriptive metadata,* in archival terms).[xlii]

Both across and, to a lesser extent, within individual collections, resources may vary in medium, form, and structure, as well as in many independent aspects of content (such as subject and genre) and context (such as place and date of production and identity of creator). In particular, some resources exist only in *analog* form, e.g., as handwritten or typewritten manuscripts, as printed publications, or as photographic prints. Others are born-digital, and remain accessible primarily in that form. Yet other resources that were originally created in analog form owe their physical accessibility (and, if described and/or available as searchable full-text, their intellectual findability) largely to their having been reproduced digitally.

Similarly, descriptive metadata may vary on several dimensions, although efforts to standardize and thus to reduce variation have had some success. Widely used national and international standards exist for:

- *data models* (sometimes also referred to as *metadata models*) that specify the kinds of entity that may be represented by metadata (viz., not only "document-like" resources, but also specific agents, events, places, etc., as well as kinds of agent, event, place, etc.), and, in the case of entity–relationship models, the relationships between those entity-types. The design of relational databases and other metadata standards may be based around such models;
- *metadata element sets* that specify the attributes of the resources of any given kind, or in any given collection, for which values may be determined and recorded;
- sets of *rules* (e.g., rules for description, or for cataloging) that provide consistent guidance, for the person who assigns metadata, in determining, for each attribute of each resource, the appropriate value, and the appropriate form in which that value is recorded;
- *controlled vocabularies* that specify the "preferred" value-types available to metadata specialists and, perhaps, to searchers; and
- the *encoding, format, and exchange* of metadata, whether as an integral component of individual resources, or in separate, independently-managed authority files.[xliii]

Just like resources, descriptive metadata may also be made available in analog form (e.g., as catalog cards) and/or in digital form (e.g., as database records). In principle, therefore, we might expect to see instances of four kinds of resource/metadata systems: (1) analog resources, analog metadata; (2) analog resources, digital metadata; (3) digital resources, analog metadata; and (4) digital resources, digital metadata. In practice, examples of systems in the third category are rare (printed directories of websites come to mind). Nonetheless, the history of the development of databases of cultural resources, such as the materials found in libraries, archives, and museums, has been one of movement, mainly in the last few decades of the twentieth century and the first few of the twenty-first, from the first of these four categories to the last. In libraries, for example, the transition from (what we might call) Phase 1 to Phase 2 began in the late 1960s, with the introduction of the MARC (Machine-Readable Cataloging) format for bibliographic records, and continued with the wholesale replacement, over the following quarter-

century, of card catalogs by OPACs (online public-access catalogs).[xliv] The move from Phase 2 to Phase 4 began in earnest with the rapid expansion of the Web in the 1990s, and latterly has been spearheaded by projects such as Google Books, the Hathi Trust Digital Library, and the Digital Public Library of America, in which massive quantities of library resources, previously available only in analog form, have been digitized and uploaded to the network for remote access online.[xlv]

5. Content, context, and structure

We alluded above to an important distinction between the *content* of a resource and the multiple *contexts* in which that resource is produced, interpreted, and used. The textual content of a published book, for example, may be distinguished from the context of its production; likewise, attributes of its content, such as subject, may be distinguished from attributes of its context, such as place of production. Furthermore, we may isolate attributes of the *structure* of a resource, such as the extent to which the component parts of its content are differentiated from one another. One textual resource may be viewed as being highly structured, in the sense that a hierarchical structure of discrete chapters, sections, and paragraphs, is clearly indicated by conventional devices in the resource itself; another may be viewed as being quite unstructured, in the sense that no such devices are used to break up a lengthy stream of text.

While they may seem facile, these distinctions are significant for discussions of IR, since the history of methods of description can be interpreted as a sequence of changes of emphasis, in each phase of which one or other of the general attribute-types—content, context, or structure—is newly highlighted (see Table 2).

Table 2. Classification of IR systems according to the kinds of attributes whose values serve as the sources of resource descriptors.

1950s-	The early development of IR systems, from the 1950s on, involved methods of automatic indexing that were based on statistical analysis, initially of the *content* of machine-readable surrogates for resources, and subsequently of the content (i.e., the "full texts") of the resources themselves.
1960s-	Later phases of development have turned the spotlight onto *context*-based approaches in which the content of resources (or records) that are *related* in some way to the original (e.g., by co-citation) is automatically analyzed in order to identify suitably discriminating descriptors.
1990s-	Link-based approaches—such as Google's assignment to every web page of a score ("PageRank"), derived from analysis of the web's structure of hypertext links among pages, that serves as an indicator of the page's relative importance within that structure—may be conceived as emphasizing the *macro-structural* attributes of collections considered as wholes.
1990s-	Yet other approaches—those developed in the burgeoning subfield of XML

	retrieval, for example—are based on the *micro-structural* analysis of the relationships among the component parts of the content of individual resources.

Overlaying this classification are several other notable distinctions and dimensions of difference (see Table 3).

Table 3. Classification of IR systems on selected dimensions.

	A	**B**
State of **resources**?	analog	digital
State of **metadata**?	analog	digital
Method of **metadata-creation**?	manual	automatic
	by assignment (potentially from other sources)	by extraction or inference (from the resource in question)
[If metadata-creation is manual:] **Selectors** of index terms?	professionals	"the crowd"
[If index terms are crowdsourced:] Method of **crowdsourcing**?	direct	indirect
State of **authority data**?	analog	digital
Authority control?	controlled	uncontrolled
Method of **authority file-creation**?	manual	automatic
	top-down	bottom-up
Interface features?	little automated search assistance	much automated search assistance
Meta-search capability?	Single repository only	multiple repositories

From this unpromisingly complex proliferation of dimensions on which IR systems may be classified (and there are surely many more), we can salvage the following

simplification—the observation of a general trend from (A) analog resources; manual metadata-creation only, by assignment, by professionals; analog controlled vocabularies, constructed manually, top-down; no automated search assistance; single-repository search; to (B) digital resources; multiple methods of manual and automatic metadata-creation in combination; digital uncontrolled "folksonomies," constructed semi-automatically, bottom-up; automated search assistance of multiple kinds; multi-repository search.

We might well be prompted to ask: How has this trajectory played out in archives? What are the determinants of the similarities and differences between the provision of access to archival resources and to resources of other kinds? What does the future hold? An essential preliminary to answering such questions is a characterization of the archival and recordkeeping information environment, and this is the subject of the following section.

The Archival Information Environment

Recordkeeping environments and associated archival traditions and practices vary considerably from country to country and from sector to sector. The archival information environment, therefore, can be either tightly or loosely bounded depending upon the conceptualization of recordkeeping and records systems that is being applied. Suffice it to say here that in some environments, archival considerations for IR will be threaded across the life of any given records system or resource and will pertain to a diversity of users, both primary (i.e., the creators and other users of active records systems) and secondary (i.e., those such as scholars, lawyers, human rights activists, hobbyists, and other members of the general public who need or wish to use resources generated by those systems, regardless of whether or not they are under archival control). In other environments, however, archival IR will encompass only activities relating to searching, retrieval, and use of resources in archival custody. In either case, resources and their associated metadata exist in a complex of ever-changing, ever-accumulating, temporally and contextually bound relationships with various other entities.[xlvi]

1. Characteristics of archives and archival resources

Archives today, whether physical or digital, may serve one or more roles. Firstly, they may serve a *recordkeeping* role as mechanisms for accountability, transparency, and institutional memory within governance, business, and other bureaucratic settings. It should be noted that in some institutional settings, especially those engaged in classified or otherwise sensitive or competitive activities, the ability to search is available only within the institution that created the records. Secondly, they may perform a *societal* or *community* role as cultural or memory institutions, which frequently places them in proximity to libraries and museums. Thirdly, and most recently, they may be engaged in *data* or *media archiving*, especially in fields such as certain sciences that generate large quantities of digital data, or in film production and preservation. What is common among almost all archives is that they typically provide access to voluminous and often unique resources whose distinctive materiality and circumstances are considered to be evidential in themselves. As already discussed, it is this evidential value of archival resources,

rather than their informational value, that tends to take priority in all archival activities. The quality of evidentiariness, and the need to locate and retrieve relevant evidence (as opposed to discrete pieces of information or data), are also what provide the most interesting challenges to traditional IR applications. There are also preservation and policy considerations that can have significant impact on the ways in which retrieval of archival content is automated. The fragility, uniqueness, and sometimes high economic value and complex ownership or confidentiality status of many archival resources can complicate the procedures by which legal, physical, and intellectual access is granted to individual items. Because of legal and security requirements, not every user is necessarily permitted the same level of access to archival resources. Data compilation and mining across online archival resources held in different archives is also increasingly of concern to personal privacy experts, and data protection laws in some countries prohibit users from compiling enough resources from different places to be able to profile subjects mentioned in those resources.

Again, as already alluded to, although digitization of archival resources is increasing the possibility of item-level or even *within-item* retrieval, the primary principle of organization in archives remains the *aggregation*—i.e., either accumulations of records generated by a single recordkeeping activity (a.k.a. a record series), or quantities of documentation created or collected by an individual or other entity (a.k.a. an archival collection). Just as the holdings of archives around the world vary in many different respects, the materials that make up a single aggregation are often highly heterogeneous in subject matter, genre, medium, form, structure, and even provenance (the principle of *respect des fonds* notwithstanding). Likewise, the component parts of a single archival item (e.g., a scrapbook containing press clippings, photographs, annotations, and greeting cards; or an electronic mail message with various kinds of attachments) can be highly heterogeneous in the same respects. These characteristics pose problems for the standardization of descriptive techniques, and reinforce the frequently expressed idea that, in general, archival resources are typically less likely to be useful in respect of their supplying *information about* a particular subject than they are in respect of their being *evidence of* a particular event or activity.

In recognition of the hierarchical structure formed (so the dominant view suggests and the descriptive standards assert[xlvii]) by the relationships obtaining among archival materials within a single aggregation, much of the description of those resources has been undertaken at levels of aggregation of broader or narrower scope (e.g., record groups, collections, fonds, series, files), rather than at the level of individual items that is the norm in libraries, or at the level of the documentary inter-relationships between fonds that the archival bond and a conceptual model such as the records continuum model would suggest.[xlviii] Unlike library catalog records, individual archival descriptions are often arranged hierarchically in the form of a finding aid or inventory, and the structure of such a description—just like the structure of any complex textual resource—can be captured in machine-readable form most consistently and effectively by encoding it using an XML-based markup standard like EAD (Encoded Archival Description).[xlix] Different kinds of metadata may be similarly hierarchically related to each other, depending upon the level of granularity and the unit of analysis (e.g., repository, fonds, series, item, digital

component). In fact, archival descriptions and recordkeeping metadata of other kinds are dynamic artifacts to which changes and additions are continually being made, as the resources that they describe become subject to new kinds of processes, interpretations, and uses, and as their physical condition deteriorates or otherwise shifts from the moment they were first conceived or created (i.e., not just after they were received by an archives and described). This expansion of the scope of metadata beyond merely the descriptive provides an exceptionally rich, but generally under-exploited infrastructure for IR.

Traditionally, the resource-discovery process in archives has been a very physically-based one. It involves first contemplating in which archives one might potentially find materials of interest; then going to any published descriptions of the holdings of those archives (e.g., finding aids); and finally, if any description appears to be promising, contacting or visiting the relevant archives to consult with a reference archivist prior to gaining physical access to the materials themselves in order to ascertain whether or not they are indeed what one wants. This situation is changing, however, as mass digitization efforts, as well as several decades of digital recordkeeping that has been generating born-digital records, are resulting in increasing quantities of archival resources being made available online. The previous lack of large quantities of digitized resources meant that techniques for automatic indexing were rarely applied systematically in the creation of metadata for archival materials. In those cases in which automatic indexing has been used to enhance retrieval of archival materials (e.g., when a search engine indexes the web pages its crawlers find), the indexing (and thus any subsequent searching) is of the content of pre-existing, machine-readable, web-accessible finding aids and/or catalog records, manually constructed as descriptions of archival resources. Reference models such as OAIS hold out the promise, not only of automated searching and retrieval of archival resources, but also of the ability to order and deliver a customized retrieval set (or dissemination information package, DIP) from a digital archive. The explicit recognition in OAIS of the role of such a customization capability validates archival concerns that IR mechanisms need to take into account particular user and resource contexts and restrictions.

Reflecting archivists' commitment to "the power of the principle of provenance,"[l] the relatively small amount of subject indexing that has been done in archives (notwithstanding a concerted push in this direction in the late 1980s and early 1990s[li]) is primarily based on analysis of data drawn from existing descriptions of the contexts in which archival materials were produced, interpreted, and used, rather than from the content of those resources. While an understanding of the provenance or resources is likely to remain stable over time, a limitation to subject indexing is that it is usually undertaken only at one particular moment, and thus reflects the perspectives of the person or institution assigning the descriptors, as well as the cultural context of that time. Archival resources, however, accumulate multiple layers of meaning and are subject to different interpretive frames over time, and these may not be supported by previously assigned descriptors.[lii] Moreover, the very few inter-processor overlap studies that have been undertaken of archival description suggest that different archivists are in any case highly unlikely to describe the same archival holdings using the same subject terms. This may partly be a result of the small amount of training in assigning subject access points

that most archivists have received, but is more probably an indication of the heterogeneity of subjects covered in many archival holdings, the different historical and contemporary expressions of and perspectives on those subjects, and difficulties in placing reasonable limits on the amount of subject indexing that an archives is able to do, to the most benefit of users.[liii]

2. Characteristics of users

Users and uses of archives remain understudied in comparison (for example) to users and uses of libraries, and it is unclear to what extent the findings of library research might be transferable into the archival domain. Archival user studies commonly draw (a) a distinction between creators, archives staff, scholar–researchers, and "the public"; (b) among scholar–researchers, a distinction between "serious researchers"—e.g., professional historians, biographers, and historic preservationists—and others; and (c) among members of the public, a distinction between those who need to access documents for personal purposes (e.g., property records, citizenship status, veteran benefits), lawyers, journalists, amateur genealogists, other avocational users, college students, and K-12 teachers and students. User studies have uncovered patterns in the ways in which archival users search for desired resources, suggesting that personal, organizational, geographical, and historical *names* of particulars (i.e., proper nouns) tend to be more popular and/or more useful as search terms than are words or phrases denoting general *concepts* or universals. The explanation usually runs as follows: Names are commonly used to identify the particular corporate bodies, persons, families, places, events, etc., that are participants in the *provenance*, or context of production, of archival resources; whereas concepts are commonly used to identify the kinds of things that are aspects of the *subjects* of archival resources. And, as already stated, archival users tend to be more interested in the evidentiary qualities of archival resources than they are in the informational qualities. So, for those who subscribe to this understanding of the primary utility of archival resources, the creation of, and provision of access to, descriptions of the *context* in which resources are produced are more important than are those of resources' *content*.

The benefits that searchers are presumed to derive from the representation, provided in finding aids, of the hierarchical, multi-level *structure* of aggregations of archival resources are, somewhat surprisingly, less well understood. The prescription of multi-level description has long been a cornerstone of archival principles and practices, and hierarchical structure is a core feature of XML-based encoding standards such as EAD and its siblings. But interface designers have struggled to translate the structural data embedded in EAD-encoded finding aids into visual displays that consistently meet users' requirements for ease of navigation. Further study of searchers' goals and preferences when navigating within and between finding aids (and between finding aids and authority data) is necessary.[liv]

Development Trajectory of Archival IR Systems

We earlier suggested that it might be useful to think of the development of IR systems in libraries as a sequence of transitions between certain phases defined in retrospect, and noted that a transition from Phase 1 (analog resources, analog metadata) to Phase 2 (analog resources, digital metadata) can be observed to have taken place largely in the 1970s and 1980s. In archives, the situation is complicated by the fact that finding aids can be considered simultaneously both as metadata (describing archival resources) and as resources; i.e., they are formulated as intellectual products, but also serve as records and thus as evidence of archives' own activities. The transition in archives from Phase 1 to Phase 2 has taken place over a longer period, and remains far from complete.[lv] At this point, however, we can begin to perceive a rough timeline of archival IR systems development that draws attention not only to recent successes but perhaps also to the kinds of advances that we might expect to be made in the near future. In particular, reference to our previous proto-classification (depicted in Table 3) leads us to the summary presented in Table 4.

Table 4. Evolution of archival IR systems.

	Up to early 1980s	**Early 1980s to present**	**Future**
State of **resources**?	analog resources	analog and growing volume of digitized and born-digital resources	digital resources
	arranged in accordance with classification, filing, and registration schemes manually applied by records creators, filing clerks, and registrars; sometimes arrangement is imposed by archivists	arranged in accordance with schemes manually or automatically applied by records creators, administrators, or software; sometimes arrangement is imposed by archivists	arranged automatically and in multiple ways
State of **metadata**?	analog descriptions: single-level catalog records (collection or item) and multi-	analog and digital descriptions: registry metadata, creator filing schemes, file transfer lists,	digital descriptions: registry metadata, creator filing schemes, single-level catalog records, multi-level finding aids, item-level descriptive

	level finding aids	single-level catalog records, multi-level finding aids, item-level descriptive metadata for individual digital resources	metadata for individual digital resources
	analog indexes to descriptions	analog and digital indexes to descriptions	digital indexes to descriptions
Method of **metadata-creation**?	manual assignment	manual assignment, and automatic inference from digital resources; end-user online tagging	primarily automatic inference; some manual assignment by creators and archivists; end-user online tagging
[If metadata-creation is manual:] **Selectors** of index terms?	creators, records administrators, archivists, and volunteers	creators, records administrators, archivists, volunteers, and end-users	creators, records administrators, archivists, volunteers, and end-users
[If index terms are crowdsourced:] Method of **crowdsourcing**?	n/a	direct and indirect	primarily indirect
State of **authority data**?	analog	analog and digital	digital
	data accessible to local institution only	some open data; most data accessible to local institution only	linked open data; some data accessible to local institution only
Authority control?	uncontrolled and controlled; emphasis on name authorities; authority forms applicable to local institution only or	controlled; increased emphasis on subject authorities; national and international standardization of	multi-tiered (local, regional, national, global) authority control; authority forms applicable to local institution only or to sector or discipline

	to sector or discipline	conceptual and data models, metadata element sets, rules for description, controlled vocabularies, encoding formats; authority forms applicable to local institution only or to sector or discipline	
Method of **authority file-creation**?	manual	manual	manual and semi-automatic
	top-down and bottom-up (creators)	top-down and bottom-up (creators)	top-down and bottom-up (creators and end-users)
	local collaboration only	local, intra- and inter-institution collaboration	local, intra- and inter-institution collaboration, global collaboration
Interface features?	no automated search assistance	EDM filing systems and records classification schemes; computer-assisted content-based keyword searching of descriptions and authority files	EDM filing systems and records classification schemes; computer-assisted content/context/structure-based search of digital resources, descriptions, and authority files; object- and pattern-matching techniques; specifications of best practices for the provision of computer-assisted access
Meta-search capability?	within-repository search	within- and cross-repository search	within-repository and universal search

1. Quasi-IR developments

Since the 1980s, progress in developing infrastructure to support archival access can be perceived as being made on five interrelated fronts that, as critical as their results are for the provision of high-performance access systems, might not strictly be considered as "IR" because they do not involve direct attention being paid to key aspects such as automated methods of indexing or relevance ranking.

In the first place, the development of national and international records management metadata and archival descriptive standards together with widespread adoption of item-level metadata standards such as METS and Dublin Core for digitized resources have brought the field closer to the goal of universal *standardization* or at least interoperability of metadata content and structure.[lvi]

Second, *cross-repository searching* is becoming increasingly easy as union databases of archival finding aids and other metadata grow larger and involve greater proportions of the institutional base in the geographical areas that they cover. OCLC's ArchiveGrid and the European Commission's Archives Portal Europe are paving the way to the construction of true archival equivalents of the library community's WorldCat, while initiatives such as the California Digital Library's Online Archive of California (OAC) and JISC's Archives Hub in the UK demonstrate what can be achieved at regional and national levels.[lvii]

Third, the magnitude of the effects on access of differences in the *user interfaces* to databases of archival collection descriptions is becoming increasingly obvious as growth is seen both in the literature on users' difficulties with interpreting online displays of finding aids, and in the number of more-or-less ad hoc trials of newly-designed interfaces being undertaken by system developers. It remains unclear to what extent, and in what respects, interface designers are taking into account the findings of user studies—partly because those findings are not always especially conclusive. There is certainly room in the future for further rigorous testing of the relationships between the presence or absence of particular features of user interfaces, and levels of different kinds of users' satisfaction with the quality of access that they experience.

Fourth, *digitization* and (where feasible) conversion to machine-readable text, of the content both of archival descriptions and of the resources-being-described themselves, are being conducted on an increasingly wide scale. Digitized materials are being linked to digital finding aids in descriptive systems and also contributed to web portals and multi-repository library, archives, and museum systems such as Europeana.[lviii]

Finally, growing effort is being expended in the construction of files of archival *authority data*, and in making sure that these authority files are used to archival information-seekers' best advantage. For example, interactive interfaces that allow users to make their own suggestions of "tags," i.e., words or phrases that are descriptive of some aspect of their experience while viewing a record or resource, have been the object of investigations into the comparative value of crowdsourcing, not only as a method of manually indexing finding aids, but also as a method of improving the richness of the lead-in (i.e., non-preferred) vocabulary in an authority file. Meanwhile, and perhaps more

significantly, several groups continue to pursue collaborative initiatives, with the aim of building shared authority files that provide national or even international control of the terms, especially of names (which are notoriously historically, culturally and politically contingent) used as access points in finding aids. In the US, the major efforts in this direction are being coordinated as components of a National Archival Authorities Infrastructure (NAAI), envisaged by Daniel Pitti (University of Virginia) and colleagues.[lix] As well as a National Archival Authorities Cooperative (NAAC), modeled on the Library of Congress's Name and Subject Authority Cooperative Programs (NACO and SACO), the NAAI vision includes global access to the authority file produced by participants in the Social Networks and Archival Context (SNAC) project.[lx] The NAAC will allow archivists from multiple participating institutions to contribute, to a shared file, authority records that comply with content and encoding standards such as ISAAR(CPF) and EAC-CPF. Meanwhile, the SNAC project is jump-starting the creation of this shared authority file through the development and use of innovative methods for the automatic extraction, from participating institutions' finding aids, of contextual metadata about the persons, corporate bodies, etc., whose names are controlled in the authority records.

Just as the library authority data in the Virtual International Authority File (VIAF) are exposed as linked open data (LOD) on the Semantic Web, each name identified by its own unique Uniform Resource Identifier (URI), the intention is for archival authority data to be accessible, readable, and actionable not only by human users, but also by web services that automate the process of establishing links among multiple, distributed sets of archival descriptions and authority records.[lxi] With the development of the Semantic Web, the promise of authority control in the archival context that was identified by David Bearman and others as early as the 1970s is finally being realized.[lxii]

2. XML retrieval

As already noted, there has been surprisingly little research in archival studies that could readily be categorized as "true" IR. Some pioneering work was undertaken in the late 1970s by Richard H. Lytle at the University of Maryland. In a limited experiment that was never replicated, Lytle compared the effectiveness of subject- and provenance-based retrieval.[lxiii] The primary body of work on archival IR, strictly defined, has been much more recent, and has focused on XML retrieval.

XML (eXtensible Markup Language) is a standard encoding format used to represent the internal structure of textual documents.[lxiv] Any resource that is "marked up" using XML takes the form of a hierarchy of nested statements about that resource's structure. Each statement at the lowest level of the hierarchy consists of a component, or *element*, of the text of the resource—a section, perhaps, or a paragraph, heading, or subheading, enclosed by a pair of labels, or *tags*, that indicate the element's type, and sometimes indicate the values of certain attributes of that element, too.

When the textual content of each resource in a collection is marked up in XML, a number of benefits accrue. For example, since structural markup is independent of presentational markup, the marked-up document can be rendered on screen in any of a number of

different styles or layouts, defined in stylesheets designed for this purpose. The information architect can then design, build, and evaluate web interfaces to collections of XML documents by tinkering with presentation styles, without necessarily having to worry about making permanent changes to documents' content or structure.

For the designers of IR systems, in particular, the prospect of having large resource collections marked up using XML is exciting because it appears to offer the opportunity to enable users, not only to identify relevant *documents*, but also to identify, as precisely as possible, the elementary *components* of each document that are most relevant to the user. Searchers might use a specialized query language like XPath or XQuery to access databases of XML-encoded resources, and to retrieve the contents only of elements that satisfy certain specified criteria.[lxv] These criteria might include not only the presence of a particular combination of keywords in an element's content (a "content-only" search), but also contextual criteria, such as the position of the element in the path that may be taken to it from the root of the tree (a "content-and-structure" search). Since queries of this type are similar to those typical of traditional database searches (cf. SQL), research in what has come to be known as XML retrieval straddles the IR and database management fields, focusing as it does on the retrieval of "semi-structured" content rather than structured (databases) or unstructured (IR).[lxvi]

Overlapping elements are not allowed in XML: in other words, the structure of an XML document must be tree-like. Given this constraint, retrieval from collections of XML-encoded resources is often viewed partly as a matter of matching (or determining the degree of similarity between) *paths*. The result of a search may be a list of elements ranked in order of their probability of relevance to the query, just as in traditional IR, but that query may well include a specification of a desired path-type, as well as of desired term-types. On the other hand, recursive structure, in which one instance of a particular element-type may be nested in another instance of the same type, is supported by XML. While this support for recursion provides great flexibility for the designer of XML schemas or document type descriptions (i.e., domain-specific statements of the structural requirements that XML documents must meet if they are to be validated as "well-formed" in their domain), it requires designers and users of search languages like XQuery to take account of a potentially bewildering set of possible scenarios.

Users of IR systems are especially likely to find an XML-search capability useful when their interests take them to areas in which lengthy documents, like most books and many archival finding aids, are the norm. Given that the archival informatics community has dedicated much time and effort to the development of an XML-based encoding standard for finding aids (i.e., EAD), it would be remarkable if archival retrieval system designers had *not* explored the potential for leveraging this existing base of ready-encoded, tree-structured resources by experimenting with XML retrieval.[lxvii] Yet, it does indeed appear that the range of options available to would-be exploiters of EAD structure remains far from exhausted. The archival literature is still replete with accounts of cases in which users are seen to have difficulties of various kinds when attempting to use online finding aids to locate resources that they want, and relatively few papers have described

implementations or (better yet) evaluations of innovative search engines in archival contexts.

One recent project that may yet inspire the further work that is much needed in this area is README (Retrieving Encoded Archival Descriptions More Effectively), carried out by a team led by Jaap Kamps at the University of Amsterdam, and reported most comprehensively in Junte Zhang's dissertation.[lxviii] Zhang's methods exemplify one of the conventional designs in IR research: construct a test collection of documents, queries, and associated relevance judgments (in this case, a set of EAD finding aids and search logs from the Dutch Nationaal Archief); build an IR system (in this case, one that is tailored for retrieval of EAD elements); and conduct ad hoc experiments with the aim of evaluating the impact, on standardized measures of recall and precision, of controlled variations in certain of the conditions under which searches take place.

Zhang's study is particularly important because it appears to be the first to test two hypotheses that previously have instead been treated in the archival literature as assumptions: (1) that the grouping of archival resources according to their *provenance* is beneficial for those seeking intellectual access to the materials; and (2) that arrangement by *original order* is similarly beneficial. These assumptions derive, of course, from two principles that have been considered central to the archival enterprise since the nineteenth century; together, they amount to the widespread conviction that it is of vital importance for an access system to take account of contextual description when determining the relevance of archival resources.

In one of his tests, Zhang compared (a) the effectiveness of retrieval when searchers used an interface that displayed retrieved EAD elements in the context of their finding aids (and that thus grouped together those elements that shared similar contexts) with (b) the retrieval effectiveness obtained when searchers used an interface that displayed retrieved elements out of context and in order of the elements' probability of relevance. His findings were that "element + provenance" ranking did indeed outperform simple "element" ranking, but (in a potentially damning result that is rather glossed over in the report) that a standard full-text retrieval system taking no account of any EAD encoding markedly outperformed both.[lxix] In another test, Zhang again compared the retrieval effectiveness of two systems: (a) in one, retrieved elements were ranked by probability of relevance, and (b) in the other, retrieved elements were listed in the original order in which they appear in each finding aid. Results indicated that relevance ranking is usually the better option.

Zhang also examined users' search behaviors, finding significant differences between inexperienced and experienced searchers of EAD finding aids, in several aspects of search activity. Despite these differences, however, the same type of system was found to work best for both user groups, indicating that efforts to personalize the archival search experience for members of different groups may not be worthwhile.

Conclusion

Research conducted by the computer scientists at the San Diego Supercomputer Center and the US National Archives at the end of the 1990s developed the Persistent Archives Technology (PAT) as an XML-based method for preserving electronic records. PAT separated document content and structure so that each could be stored separately in software-independent form in a way that the document could subsequently be reconstructed. To do this, the document structures of electronic records were computationally inferred from commonalities in structure across similar types of documents, and an XML DTD created on the fly for those structures.[lxx] This process had two interesting implications for IR. First, it facilitated searching and retrieval according to document structure, rather than by content or language. Second, it allowed researchers to identify documents with anomalous structures, i.e., documents that for some reason were not similar to others within the same aggregation.

The latter is a particularly interesting approach for archival IR that bears further investigation, since it suggests a potential strategy for helping users such as historians and lawyers who may be hoping to find previously unknown, and possibly "smoking gun"–type documents. It also suggests, conversely, the application of archival IR in efforts to establish the absence (as opposed to the presence) of documents, and thus to meet the archivist's goal of ensuring that records-creators are seen to be held accountable for their actions. In general, ideas such as these point to ways in which advances in archival IR that exploits multiple types and sources of metadata may find wider application in other domains where similarly rich contextual metadata exists, e.g., litigation support systems, news retrieval, audiovisual archives, data mining, and digital asset management.

It seems highly likely that, at least in the short term, XML retrieval will continue as the most productive source of inspiration for archival IR system design. On a final note: the Initiative for the Evaluation of XML Retrieval (INEX) was established in 2002 as a venue for researchers to compare their IR systems' performance in a variety of controlled environments.[lxxi] Recent tracks at INEX have focused on linked data, tweet contextualization, snippet retrieval, and social book search. Finding-aid element search may not (yet?) be the highest of INEX participants' priorities, but those interested in contributing to the development of the next generation of archival IR systems could do worse than engage in a concentrated study of the results presented annually at INEX workshops.

References

Akmon, Dharma, Ann Zimmerman, Daniels, Morgan Daniels, and Margaret Hedstrom. "The Application of Archival Concepts to a Data-intensive Environment: Working with Scientists to Understand Data Management and Preservation Needs." *Archival Science* 11, nos. 3–4 (2011): 329–348.

Archival Education and Research Institute (AERI) Pluralizing the Archival Curriculum Group (PACG). "Educating for the Archival Multiverse." *American Archivist* 74, no. 1 (2011): 69–101.

Avram, Henriette D., Ruth S. Freitag, and Kay D. Guiles. *A Proposed Format for a Standardized Machine-readable Catalog Record*. Washington, DC: Library of Congress, 1965.

Baca, Murtha, ed. *Introduction to Metadata*. 2nd ed. Los Angeles: Getty Research Institute, 2008. http://www.getty.edu/research/publications/electronic_publications/intrometadata/index.html.

Baeza-Yates, Ricardo, and Berthier Ribeiro-Neto. *Modern Information Retrieval: The Concepts and Technology behind Search*. 2nd ed. Harlow, UK: Addison-Wesley, 2011.

Bearman, David A. "Authority Control Issues and Prospects." *American Archivist* 52, no. 3 (1989): 286–299.

Bearman, David A. "Automated Access to Archival Information: Assessing Systems." *American Archivist* 42, no. 2 (1979): 179–190.

Bearman, David A., and Richard H. Lytle. "The Power of the Principle of Provenance." *Archivaria* 21 (Winter 1985–86): 14–27.

Bountouri, Lina, and Manolis Gergatsoulis. "The Semantic Mapping of Archival Metadata to the CIDOC CRM Ontology." *Journal of Archival Organization* 9, nos. 3–4 (2011): 174–207.

Buckland, Michael K. "What is a 'Document'?" *Journal of the American Society for Information Science* 48, no. 9 (1998): 804–809.

Cleverdon, Cyril W. "The Cranfield Tests on Index Language Devices." *Aslib Proceedings* 19, no. 6 (1967): 173–194.

Cleverdon, Cyril W. "The Significance of the Cranfield Tests on Index Languages." In *SIGIR '91: Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3–12. New York: ACM Press, 1991.

Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS)*. Recommended Practice CCSDS 650.0-M-2. 2012. http://public.ccsds.org/publications/archive/650x0m2.pdf.

Croft, W. Bruce, Donald Metzler, and Trevor Strohman. *Search Engines: Information Retrieval in Practice*. Boston, MA: Addison-Wesley, 2010.

Daines, J. Gordon, III, and Cory L. Nimer. "Re-imagining Archival Display: Creating User-friendly Finding Aids." *Journal of Archival Organization* 9, no. 1 (2011): 4–31.

Dooley, Jackie M. “Subject Indexing in Context.” *American Archivist* 55, no. 2 (1992): 344–354.

Dooley, Jackie M., and Helena Zinkham. “The Object as ‘Subject’: Providing Access to Genres, Forms of Material, and Physical Characteristics.” In *Beyond the Book: Extending MARC for Subject Access,* edited by Toni Peterson and Pat Molholt, 43–80. Boston: G. K. Hall, 1990.

Duranti, Luciana. “The Archival Bond.” *Archives and Museum Informatics* 11, nos. 3–4 (1997): 213–218.

Elings, Mary W., and Günter Waibel. “Metadata for All: Descriptive Standards and Metadata Sharing Across Libraries, Archives, and Museums.” *First Monday* 12, no. 3 (2007). http://firstmonday.org/ojs/index.php/fm/article/viewArticle/1628.

Evans, Frank B., Donald F. Harrison, Edwin A. Thompson, and William L. Rofes. “A Basic Glossary for Archivists, Manuscript Curators, and Records Managers.” *American Archivist* 37, no. 3 (1974): 415–433.

Frusciano, Thomas J. “Online Finding Aids, Catalog Records, and Access—Revisited.” *Journal of Archival Organization* 9 no. 1 (2011): 1–3.

Furner, Jonathan. “On Recommending.” *Journal of the American Society for Information Science and Technology* 53, no. 9 (2002): 747–763.

Furner, Jonathan. “Folksonomies.” In *Encyclopedia of Library and Information Sciences*. 3rd ed. Edited by Marcia J. Bates and Mary Niles Maack, 1858–1866. Boca Raton, FL: CRC Press, 2010.

Gilliland, Anne J. “Setting the Stage.” In *Introduction to Metadata*. 2nd ed. Edited by Murtha Baca, 1–19. Los Angeles: Getty Research Institute, 2008. http://www.getty.edu/research/publications/electronic_publications/intrometadata/setting.html.

Gilliland, Anne J. *Conceptualizing Twenty-first-century Archives*. Chicago, IL: Society of American Archivists, in press.

Greene, Mark A. “MPLP: It’s Not Just for Processing Any More,” *American Archivist* 73, no. 1 (2010): 175–203.

Harman, Donna. *Information Retrieval Evaluation*. San Rafael, CA: Morgan & Claypool, 2011.

Harpring, Patricia. *Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works.* Los Angeles: Getty Research Institute, 2010.

http://www.getty.edu/research/publications/electronic_publications/intro_controlled_vocab/.

Hopper, Grace M. "A Glossary of Computer Terminology." *Computers and Automation* 3, no. 5 (1954): 14–18, 20, 22.

International Council on Archives. *General International Standard Archival Description*. 2nd ed. Paris: International Council on Archives, 1999.

International Council on Archives. *International Standard Archival Authority Record for Corporate Bodies, Persons, and Families*. 2nd ed. Paris: International Council on Archives, 2004.

International Organization for Standardization. *ISO/IEC 2382-4:1999 Information Technology — Vocabulary — Part 4: Organization of Data.* Geneva, Switzerland: ISO, 1999.

International Organization for Standardization. *ISO 5127:2001 Information and Documentation — Vocabulary.* Geneva, Switzerland: ISO, 2001.

International Organization for Standardization. *ISO 8459:2009 Information and Documentation — Bibliographic Data Element Directory for Use in Data Exchange and Inquiry.* Geneva, Switzerland: ISO, 2009.

International Organization for Standardization. *ISO 14721:2012 Space Data and Information Transfer Systems — Open Archival Information System (OAIS) — Reference Model*. Geneva, Switzerland: ISO, 2012.

International Organization for Standardization. *ISO 15836:2009 Information and Documentation — The Dublin Core Metadata Element set*. Geneva, Switzerland: ISO, 2009.

International Organization for Standardization. *ISO/IEC/IEEE 24765:2010 Systems and Software Engineering — Vocabulary.* Geneva, Switzerland: ISO, 2010.

International Organization for Standardization. *ISO 25964-1:2011 Information and Documentation — Thesauri and Interoperability with Other Vocabularies — Part 1: Thesauri for Information Retrieval.* Geneva, Switzerland: ISO, 2011.

International Organization for Standardization. *ISO 30300:2011 Information and Documentation — Management Systems for Records — Fundamentals and Vocabulary.* Geneva, Switzerland: ISO, 2011.

International Organization for Standardization, Technical Committee 46, Subcommittee 11 (ISO/TC 46/SC 11): Archives/Records Management. *Relationship between the ISO 30300 Series of Standards and Other Products of ISO/TC 46/SC 11:2. Vocabulary.* 2012.

http://www.niso.org/apps/group_public/download.php/9745/White_paper-Rel.ship_30300_standards-VOCABULARY-v5.pdf.

Lalmas, Mounia. *XML Retrieval*. San Rafael, CA: Morgan & Claypool, 2009.

Lalmas, Mounia, and Anastasios Tombros. "Evaluating XML Retrieval Effectiveness at INEX." *ACM SIGIR Forum* 41, no. 1 (2007): 40–57.

Lancaster, F. Wilfred. *Information Retrieval Systems; Characteristics, Testing, and Evaluation.* New York: Wiley, 1968.

Ludaescher, Bertram, Richard Marciano, and Reagan Moore. "Towards Self-validating Knowledge-based Archives." In *11th Workshop on Research Issues in Data Engineering.* Heidelberg, Germany: IEEE Computer Society, 2001. http://www.sdsc.edu/~ludaesch/Paper/ride01.html.

Lytle, Richard H. "Subject Retrieval in Archives: A Comparison of the Provenance and Content Indexing Methods." PhD diss., University of Maryland, 1979.

Lytle, Richard H. "Intellectual Access to Archives: 1. Provenance and Content Indexing Methods of Subject Retrieval." *American Archivist* 43, no. 1 (1980): 64–75.

Lytle, Richard H. "Intellectual Access to Archives: II. Report of an Experiment Comparing Provenance and Content Indexing Methods of Subject Retrieval." *American Archivist* 43, no. 2 (1980): 191–207.

Maier, Shannon Bowen. "MPLP and the Catalog Record as a Finding Aid." *Journal of Archival Organization* 9 no. 1 (2011): 32–44.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. New York, NY: Cambridge University Press, 2008.

Mascaro, Michelle. "Controlled Access Headings in EAD Finding Aids: Current Practices in Number of and Types of Headings Assigned." *Journal of Archival Organization* 9, nos. 3–4 (2011): 208–225.

McCallum, Sally H. "An Introduction to the Metadata Object Description Schema (MODS)." *Library Hi-Tech* 22, no. 1 (2004): 82–88.

McKemmish, Sue. "Traces: Document, Record, Archive, Archives." In *Archives: Recordkeeping in Society*, edited by Sue McKemmish, Michael, Piggott, Barbara Reed, and Frank Upward. Wagga Wagga, Australia: Centre for Information Studies, Charles Sturt University, 2005.

McKemmish, Sue, Glenda Acland, Nigel Ward, and Barbara Reed. “Describing Records in Context in the Continuum: The Australian Recordkeeping Metadata Schema.” *Archivaria* 48 (Fall 1999): 3–42.

Michelson, Avra. “Description and Reference in the Age of Automation.” *American Archivist* 50, no. 2 (1987): 192–208.

Miller, Eric. “An Introduction to the Resource Description Framework.” *D-Lib Magazine* 4, no. 5 (1998). http://www.dlib.org/dlib/may98/miller/05miller.html.

Mooers, Calvin N. “Coding, Information Retrieval, and the Rapid Selector.” *American Documentation* 1, no. 4 (1950): 225–229.

Moore, Reagan, Chaitan Baru, Arcot Rajasekar, Bertram Ludaescher, Richard Marciano, Michael Wan, Wayne Schroeder, and Amarnath Gupta. “Collection-based Persistent Digital Archives: Part 1.” *D-Lib Magazine* 6, no. 3 (2000). http://www.dlib.org/dlib/march00/moore/03moore-pt1.html.

Moore, Reagan, Chaitan Baru, Arcot Rajasekar, Bertram Ludaescher, Richard Marciano, Michael Wan, Wayne Schroeder, and Amarnath Gupta. “Collection-based Persistent Digital Archives: Part 2.” *D-Lib Magazine* 6, no. 4 (2000). http://www.dlib.org/dlib/april00/moore/04moore-pt2.html.

Ostroff, Harriet. “Subject Access to Archival and Manuscript Material.” *American Archivist* 53, no. 1 (1990): 100–105.

Panizzi, Antonio. “Rules for the Compilation of the Catalogue.” In *Catalogue of Printed Books in the British Museum*, Vol. 1, v–ix. London: Trustees of the British Museum, 1841.

Pearce-Moses, Richard. *A Glossary of Archival & Records Terminology.* Chicago: Society of American Archivists, 2005.

Pitti, Daniel V. “National Archival Authorities Infrastructure.” Accessed April 28, 2013. http://ecommons.cornell.edu/bitstream/1813/28718/7/Pitti_SNAC-NAAC_Cornell.pdf.

Pugh, Mary Jo. “The Illusion of Omniscience: Subject Access and the Reference Archivist.” *American Archivist* 45, no. 1 (1982): 35–36.

Roelleke, Thomas. *Information Retrieval Models: Foundations and Relationships*. San Rafael, CA: Morgan & Claypool, 2013.

Salton, Gerard. “A Comparison Between Manual and Automatic Indexing Methods.” *American Documentation* 20, no. 1 (1969): 61–71.

Salton, Gerard, and Christopher Buckley. “Term Weighting Approaches in Automatic Text Retrieval.” *Information Processing & Management* 24, no. 5 (1988): 513–523.

Sanderson, Mark and W. Bruce Croft. “A History of Information Retrieval Research.” *Proceedings of the IEEE* 100 (May 13, 2012): 1444–1451.

Schwartz, Michael F. “The Networked Resource Discovery Project.” In *Proceedings of the IFIP XI World Congress*, 827–832. 1989.

Smiraglia, Richard. “Subject Access to Archival Materials Using LCSH.” *Cataloging & Classification Quarterly* 11, nos. 3–4 (1990): 63–90.

Society of American Archivists. *Describing Archives: A Content Standard*. 2nd ed. Chicago: Society of American Archivists, 2013.

Spärck Jones, Karen. “The Cranfield Tests.” In *Information Retrieval Experiment*, edited by Karen Spärck Jones, 256–284. London: Butterworths, 1981.

Spärck Jones, Karen. “Retrieval System Tests 1958–1978.” In *Information Retrieval Experiment*, edited by Karen Spärck Jones, 213–255. London: Butterworths, 1981.

Spärck Jones, Karen. “Reflections on TREC.” *Information Processing & Management* 31, no. 3 (1995): 291–314.

Spärck Jones, Karen. “Further Reflections on TREC.” *Information Processing & Management* 36, no. 1 (2000): 37–85; and

Spärck Jones, Karen. “What’s the Value of TREC?” *ACM SIGIR Forum* 40, no. 1 (2006): 10–20.

Upward, Frank, Sue McKemmish, and Barbara Reed. “Archivists and Changing Social and Information Spaces: A Continuum Approach to Recordkeeping and Archiving in Online Cultures.” *Archivaria* 72 (Fall 2011): 197–238.

Walne, Peter, ed. *Dictionary of Archival Terminology: English and French; with Equivalents in Dutch, German, Italian, Russian and Spanish*. 2nd rev. ed. München: K. G. Saur, 1988.

Weibel, Stuart. “Metadata: The Foundations of Resource Description.” *D-Lib Magazine* 1, no. 1 (1995). http://www.dlib.org/dlib/July95/07weibel.html.

White, Kelvin L., and Anne J. Gilliland. “Promoting Reflexivity and Inclusivity in Archival Education, Research, and Practice.” *Library Quarterly* 80, no. 3 (2010): 231–248.

Wilmot, Erroll de Burgh, ed. *Glossary of Terms Used in Automatic Data Processing*. London: Business Publications Limited, 1960.

Wilson, Max L. *Search User Interface Design*. San Rafael, CA: Morgan & Claypool, 2012.

Woodley, Mary. *DCMI Glossary*. 2005. http://dublincore.org/documents/usageguide/glossary.shtml.

Zhang, Jane. "Archival Representation in the Digital Age." *Journal of Archival Organization* 10, no. 1 (2012): 45–68.

Zhang, Junte. "System Evaluation of Archival Description and Access." PhD diss., University of Amsterdam, 2011. http://www.illc.uva.nl/Research/Dissertations/DS-2011-04.text.pdf.

[i] See, e.g., Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, *Introduction to Information Retrieval* (New York, NY: Cambridge University Press, 2008); Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology behind Search*, 2nd ed. (Harlow, UK: Addison-Wesley, 2011).
[ii] See, e.g., W. Bruce Croft, Donald Metzler, and Trevor Strohman, *Search Engines: Information Retrieval in Practice* (Boston, MA: Addison-Wesley, 2010).
[iii] See, e.g., Max L. Wilson, *Search User Interface Design* (San Rafael, CA: Morgan & Claypool, 2012).
[iv] See, e.g., Donna Harman, *Information Retrieval Evaluation* (San Rafael, CA: Morgan & Claypool, 2011).
[v] See, e.g., Thomas Roelleke, *Information Retrieval Models: Foundations and Relationships* (San Rafael, CA: Morgan & Claypool, 2013).
[vi] David Bearman was probably the first to use the term "informatics" in an archival context. Bearman's consulting firm, Archives & Museum Informatics, began publishing the *Archival Informatics Newsletter* in 1987; this journal was itself retitled *Archives and Museum Informatics* in 1989, and *Archival Science* in 2000. The principal catalyst for debates about the value of subject indexing in archival settings was Richard H. Lytle's doctoral dissertation of 1979, "Subject Retrieval in Archives: A Comparison of the Provenance and Content Indexing Methods" (University of Maryland).
[vii] We use the term "archival studies" here because it encompasses "the fullest range of archival practice, ideas, and research from multiple professional, community, and disciplinary perspectives"; see Archival Education and Research Institute (AERI) Pluralizing the Archival Curriculum Group (PACG), "Educating for the Archival Multiverse," *American Archivist* 74, no. 1 (2011): 72. See also Kelvin L. White and Anne J. Gilliland, "Promoting Reflexivity and Inclusivity in Archival Education, Research, and Practice," *Library Quarterly* 80, no. 3 (2010): 231–248.
[viii] See, e.g., Sue McKemmish, Glenda Acland, Nigel Ward, and Barbara Reed, "Describing Records in Context in the Continuum: The Australian Recordkeeping Metadata Schema," *Archivaria* 48 (Fall 1999): 3–37; and Dharma Akmon, Ann Zimmerman, Morgan Daniels, and Margaret Hedstrom, "The Application of Archival Concepts to a Data-intensive Environment: Working with Scientists to Understand Data Management and Preservation Needs," *Archival Science* 11, nos. 3–4 (2011): 329–348.
[ix] See, e.g., Lina Bountouri and Manolis Gergatsoulis, "The Semantic Mapping of Archival Metadata to the CIDOC CRM Ontology," *Journal of Archival Organization* 9, nos. 3–4 (2011): 174–207; Sally H. McCallum, "An Introduction to the Metadata Object Description Schema (MODS)," *Library Hi-Tech* 22, no. 1 (2004): 82–88; Michelle Mascaro, "Controlled Access Headings in EAD Finding Aids: Current Practices in Number of and Types of Headings Assigned," *Journal of Archival Organization* 9, nos.3–4 (2011): 208–225; Thomas J. Frusciano, "Online Finding Aids, Catalog Records, and Access—Revisited," *Journal of Archival Organization* 9, no. 1 (2011): 1–3; Jane Zhang, "Archival Representation in the Digital Age," *Journal of Archival Organization* 10, no. 1 (2012): 45–68; Shannon Bowen Maier, "MPLP and the Catalog Record as a Finding Aid," *Journal of Archival Organization* 9, no. 1 (2011): 32–44 ; and Mark A. Greene, "MPLP: It's Not Just for Processing Anymore," *American Archivist* 73, no. 1 (2010): 175–203.
[x] In 1937, the National Archives was a founding member of the American Documentation Institute (ADI). The forerunner of today's Association for Information Science and Technology (ASIS&T), ADI was concerned with the classification of, and access to, scientific and social scientific documentation. This involvement ceased when the archives was subsumed into the General Services Administration in 1941, and the archival field in the US did not continue to maintain any kind of close relationship with the information science field that was so instrumental in the development of IR.
[xi] See, e.g., Michael K. Buckland, "What is a 'Document'?" *Journal of the American Society for Information Science* 48, no. 9 (1998): 804–809.
[xii] Cyril W. Cleverdon, "The Cranfield Tests on Index Language Devices," *Aslib Proceedings* 19, no. 6 (1967): 173–194; F. Wilfred Lancaster, *Information Retrieval Systems: Characteristics, Testing, and Evaluation* (New York: Wiley, 1968).
[xiii] See, e.g., Calvin N. Mooers, "Coding, Information Retrieval, and the Rapid Selector," *American Documentation* 1, no. 4 (1950): 225–229; see also Mark Sanderson and W. Bruce Croft, "A History of Information Retrieval Research," *Proceedings of the IEEE* 100 (May 13, 2012): 1444–1451.

[xiv] In some archival contexts, "records" is reserved specifically to refer to organizational, business, governmental, public, or legal records, to be contrasted with the "papers" or "manuscripts" that provide evidence of the activities of individual persons and families.

[xv] *ISO 30300:2011 Information and Documentation — Management Systems for Records — Fundamentals and Vocabulary* is the most recent of a sequence of ISO standards defining "record(s)" as "information created, received, and maintained as evidence and information by an organization or person, in pursuit of legal obligations or in the transaction of business"—a definition with clear echoes both of that supplied by Frank B. Evans, Donald F. Harrison, Edwin A. Thompson, and William L. Rofes in "A Basic Glossary for Archivists, Manuscript Curators, and Records Managers," *American Archivist* 37, no. 3 (1974): 415–433 ("all recorded information, regardless of media or characteristics, made or received and maintained by an organization or institution in pursuance of its legal obligations or in the transaction of its business"), and of the one given in the U.S. Government's *Records Disposal Act* of 1943 ("all … documentary materials, regardless of physical form or characteristics, made or received … in pursuance of Federal law or in connection with the transaction of public business, and preserved … as evidence … or because of the informational value of data contained therein").

[xvi] *ISO/IEC/IEEE 24765:2010 Systems and Software Engineering — Vocabulary* defines "record" as "a set of related data items treated as a unit"; *ISO/IEC 2382-4:1999 Information Technology — Vocabulary — Part 4: Organization of Data* talks of "elements" instead of "items." One of the earliest dictionary definitions of this sense of "records" appears in *Glossary of Terms Used in Automatic Data Processing*, ed. Erroll de Burgh Wilmot (London: Business Publications Limited, 1960): "items of information constituting a complete file." Since at least 1954, "item" had been used for this purpose: see, e.g., Grace M. Hopper, "A Glossary of Computer Terminology," *Computers and Automation* 3, no. 5 (1954): 14–18, 20, 22, where "item" is defined as "a set of one or more fields containing related information."

[xvii] Although *ISO 8459:2009 Information and Documentation — Bibliographic Data Element Directory for Use in Data Exchange and Inquiry* defines "record" similarly to *ISO/IEC/IEEE 24765:2010*, as "group of data elements usually treated as a unit and often organized into sub-units called fields, which identifies, describes, and facilitates retrieval of an entity," it also defines "catalogue record" as "record in a cataloguing system that describes, analyses, and controls bibliographic, authority, or holdings data." The use of "entry" with this sense may be found in Antonio Panizzi's "Rules for the Compilation of the Catalogue," in *Catalogue of Printed Books in the British Museum*, Vol. 1 (London: Trustees of the British Museum, 1841), v–ix. The term "catalog record" gained currency with the first reports of the Library of Congress's Machine-Readable Cataloging (MARC) Project in the mid-1960s—see, e.g., Henriette D. Avram, Ruth S. Freitag, and Kay D. Guiles, *A Proposed Format for a Standardized Machine-readable Catalog Record* (Washington, DC: Library of Congress, 1965)—but the "Glossary" appearing as Appendix D in the *Anglo-American Cataloguing Rules*, 2nd ed. (Chicago: American Library Association, 1978), 563–572, still defines "entry" as "a record of an item in a catalogue," without also defining "record." The widespread use of "record" in preference to "entry" in the library context dates rather from the 1980s, with the sharp acceleration during that period in a shift from card catalogs to OPACs (online public access catalogs).

[xviii] Sense 2 of "collection" in *ISO 5127:2001 Information and Documentation — Vocabulary*.

[xix] See, e.g., *Describing Archives: A Content Standard*, 2nd ed. (Chicago: Society of American Archivists, 2013), 21.

[xx] Synonymous, in other words, with the terms "fonds" and "record group": see, e.g., the "Notes" to the entry for "Collection" in Richard Pearce-Moses, *A Glossary of Archival & Records Terminology* (Chicago: Society of American Archivists, 2005), 76.

[xxi] See, e.g., sense 3 of "collection" in *ISO 5127:2001 Information and Documentation — Vocabulary*.

[xxii] Richard Pearce-Moses, *A Glossary of Archival & Records Terminology* (Chicago: Society of American Archivists, 2005), 244.

[xxiii] One of the pioneer users of "resource" in this way was Michael F. Schwartz—see, e.g., his "The Networked Resource Discovery Project," in *Proceedings of the IFIP XI World Congress* (1989), 827–832. In the mid-1990s, the Online Computer Library Center (OCLC) embarked on an Internet Resource Cataloging Project, leading to a Metadata Workshop in Dublin, OH, in March 1995, convened in collaboration with the National Center for Supercomputing Applications—see, e.g., Stuart Weibel, "Metadata: The Foundations of Resource Description," *D-Lib Magazine* 1, no. 1 (1995), http://www.dlib.org/dlib/July95/07weibel.html. The outcome of this work, critical to the coalescence of the

digital library field, was the Dublin Core Metadata Element Set, a specification of "a core set of metadata elements to describe networked resources" originally codified in *RFC 2413:1998 Dublin Core Metadata for Resource Discovery* and most recently standardized as *ISO 15836:2009 Information and Documentation — The Dublin Core Metadata Element Set*. "Resource" is defined in *ISO 15836* as "anything that may be identified," and in the World Wide Web Consortium's specification of the Resource Description Framework (RDF) as "any object that is uniquely identifiable by a Uniform Resource Identifier (URI)"—see, e.g., Eric Miller, "An Introduction to the Resource Description Framework," *D-Lib Magazine* 4, no. 5 (1998), http://www.dlib.org/dlib/may98/miller/05miller.html. The library community's *RDA: Resource Description and Access*, published in 2010 as a replacement for the *Anglo-American Cataloging Rules*, 2nd ed. (AACR2), defines "resource" more narrowly as "a work, expression, manifestation, or item," thus excluding entities associated with those resources such as persons, corporate bodies, families, concepts, objects, events, and places; see http://www.rdatoolkit.org/ for more on RDA.

[xxiv] International Organization for Standardization, Technical Committee 46, Subcommittee 11 (ISO/TC 46/SC 11): Archives/Records Management, *Relationship between the ISO 30300 Series of Standards and Other Products of ISO/TC 46/SC 11:2. Vocabulary* (2012), http://www.niso.org/apps/group_public/download.php/9745/White_paper-Rel.ship_30300_standards-VOCABULARY-v5.pdf.

[xxv] See http://www.icacds.org.uk/eng/standards.htm for more on ISAD.

[xxvi] The most recent ICA-approved source is *Dictionary of Archival Terminology: English and French; with Equivalents in Dutch, German, Italian, Russian and Spanish*, ed. Peter Walne, 2nd rev. ed. (München: K. G. Saur, 1988).

[xxvii] Mary Woodley, *DCMI Glossary* (2005), http://dublincore.org/documents/usageguide/glossary.shtml.

[xxviii] ISO/TC 46/SC 11, *Relationship between the ISO 30300 Series of Standards and Other Products*.

[xxix] The choice of terminology in *ISO 14721:2012* is justified as follows: "As this reference model is applicable to all disciplines and organizations that do, or expect to, preserve and provide information in digital form, these terms cannot match all of those familiar to any particular discipline (e.g., traditional Archives, digital libraries, science data centers). Rather, the approach taken is to use terms that are not already overloaded with meaning so as to reduce conveying unintended meanings. Therefore it is expected that all disciplines and organizations will find that they need to map some of their more familiar terms to those of the OAIS Reference Model. This should not be difficult and is viewed as a contribution, rather than a deterrent, to the success of the reference model. For example, archival science focuses on preservation of the 'record'. This term is not used in the OAIS Reference Model, but one mapping might approximately equate it with 'Content Information within an Archival Information Package'." See Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS)*, Recommended Practice CCSDS 650.0-M-2 (2012), http://public.ccsds.org/publications/archive/650x0m2.pdf.

[xxx] *ISO 14721:2012* uses the term "archive" not to refer to resources, but to a kind of organization; see "7. Resource-management organizations and systems," below in Table 1.

[xxxi] See, e.g., Murtha Baca, ed., *Introduction to Metadata*, 2nd ed. (Los Angeles: Getty Research Institute, 2008), http://www.getty.edu/research/publications/electronic_publications/intrometadata/index.html.

[xxxii] See, e.g., Anne J. Gilliland, "Setting the Stage," in *Introduction to Metadata*, 2nd ed., ed. Murtha Baca (Los Angeles: Getty Research Institute, 2008), 1–19, http://www.getty.edu/research/publications/electronic_publications/intrometadata/setting.html.

[xxxiii] See, e.g., Gerard Salton, "A Comparison Between Manual and Automatic Indexing Methods," *American Documentation* 20, no. 1 (1969): 61–71.

[xxxiv] See Cyril W. Cleverdon, "The Cranfield Tests on Index Language Devices," *Aslib Proceedings* 19, no. 6 (1967): 173–194. See also Karen Spärck Jones, "The Cranfield Tests," in *Information Retrieval Experiment*, ed. Karen Spärck Jones (London: Butterworths, 1981), 256–284; Cyril W. Cleverdon, "The Significance of the Cranfield Tests on Index Languages," in *SIGIR '91: Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York: ACM Press, 1991), 3–12.

[xxxv] See, e.g., Karen Spärck Jones, "Retrieval System Tests 1958–1978," in *Information Retrieval Experiment*, ed. Karen Spärck Jones (London: Butterworths, 1981), 213–255; and Gerard Salton and Christopher Buckley, "Term Weighting Approaches in Automatic Text Retrieval," *Information Processing & Management* 24, no. 5 (1988): 513–523. See also Karen Spärck Jones, "Reflections on TREC,"

Information Processing & Management 31, no. 3 (1995): 291–314; "Further Reflections on TREC," *Information Processing & Management* 36, no. 1 (2000): 37–85; and "What's the Value of TREC?" *ACM SIGIR Forum* 40, no. 1 (2006): 10–20.

[xxxvi] See http://id.loc.gov/authorities/subjects.html for more on LCSH.

[xxxvii] See http://www2.archivists.org/standards/describing-archives-a-content-standard-second-edition-dacs for more on DACS; http://www.ica.org/?lid=10203 for ISAAR(CPF); and http://eac.staatsbibliothek-berlin.de/ for EAC-CPF.

[xxxviii] See, e.g., Patricia Harpring, *Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works* (Los Angeles: Getty Research Institute, 2010), http://www.getty.edu/research/publications/electronic_publications/intro_controlled_vocab/.

[xxxix] See http://www.ica.org/?lid=10208 for ISDF.

[xl] See, e.g., Jonathan Furner, "Folksonomies," in *Encyclopedia of Library and Information Sciences*, 3rd ed., ed. Marcia J. Bates and Mary Niles Maack (Boca Raton, FL: CRC Press, 2010), 1858–1866.

[xli] See, e.g., Jonathan Furner, "On Recommending," *Journal of the American Society for Information Science and Technology* 53, no. 9 (2002): 747–763. For those who are uncomfortable with the suggestion that recommender systems are based, even if "indirectly," on manual indexing, a slightly more attractive alternative might be to view such systems as "semi-automatic," reflecting the sense in which people, rather than machines, are the ultimate executors of the acts, representations of which are used to characterize individual resources.

[xlii] It should be obvious that, here, we are choosing not to use "record" in the sense of "catalog record," "bibliographic record," or even "metadata record," in order (a) to avoid confusion with the senses in which it is used in archives and recordkeeping, and (b) to reflect the increasing tendency, even within library contexts, to emphasize data and metadata rather than records.

[xliii] See, e.g., Mary W. Elings and Günter Waibel, "Metadata for All: Descriptive Standards and Metadata Sharing Across Libraries, Archives, and Museums," *First Monday* 12, no. 3 (2007), http://firstmonday.org/ojs/index.php/fm/article/viewArticle/1628.

[xliv] See http://www.loc.gov/marc/ for more on MARC.

[xlv] See http://books.google.com/ for Google Books, http://www.hathitrust.org/ for the HathiTrust Digital Library, and http://dp.la/ for the Digital Public Library of America.

[xlvi] The existence of this complex is an underlying assumption of the ISO Records Management standards and is specifically delineated in the records continuum model and the entity–relationship (ER) conceptual model for the Australian Recordkeeping Metadata Schema (RKMS) that has been the basis for the ER conceptual model currently being developed for ICA archival descriptive standards. RKMS specifies multiple entities, including agents, mandates and functions. See McKemmish, et al. "Describing Records in Context in the Continuum."

[xlvii] See International Council on Archives, *General International Standard Archival Description*, 2nd ed. (Paris: International Council on Archives, 1999).

[xlviii] See Luciana Duranti, "The Archival Bond," *Archives and Museum Informatics* 11, nos. 3–4 (1997): 213–218; Frank Upward, Sue McKemmish and Barbara Reed, "Archivists and Changing Social and Information Spaces: A Continuum Approach to Recordkeeping and Archiving in Online Cultures," *Archivaria* 72 (Fall 2011): 197–238; Sue McKemmish, "Traces: Document, Record, Archive, Archives," in Sue McKemmish, Michael Piggott, Barbara Reed, and Frank Upward, eds., *Archives: Recordkeeping in Society* (Wagga Wagga, Australia: Centre for Information Studies, Charles Sturt University, 2005); McKemmish et al. "Describing Records in Context in the Continuum."

[xlix] See http://www.loc.gov/ead/ for more on EAD.

[l] This is the title of an influential paper by David A. Bearman and Richard H. Lytle, *Archivaria* 21 (Winter 1985–86): 14–27.

[li] Earlier studies and position papers relating to enhancing subject access to archives include Mary Jo Pugh, "The Illusion of Omniscience: Subject Access and the Reference Archivist," *American Archivist* 45, no. 1 (1982): 35–36; Avra Michelson, "Description and Reference in the Age of Automation," *American Archivist* 50, no. 2 (1987): 192–208; David Bearman, "Authority Control Issues and Prospects," *American Archivist* 52, no. 3 (1989): 286–299; Jackie M. Dooley and Helena Zinkham, "The Object as 'Subject': Providing Access to Genres, Forms of Material, and Physical Characteristics," in *Beyond the Book: Extending MARC for Subject Access,* ed. Toni Peterson and Pat Molholt (Boston: G. K. Hall, 1990), 43–80; Harriet Ostroff, "Subject Access to Archival and Manuscript Material," *American Archivist* 53, no. 1

(1990): 100–105; Richard Smiraglia, "Subject Access to Archival Materials Using LCSH," *Cataloging & Classification Quarterly* 11, nos. 3–4 (1990): 63–90; and Jackie M. Dooley, "Subject Indexing in Context," *American Archivist* 55, no. 2 (1992): 344–354.

[lii] In some contexts, social tagging might offer mitigation of this concern.

[liii] See Michelson, "Description and Reference in the Age of Automation."

[liv] See, e.g., J. Gordon Daines, III, and Cory L. Nimer, "Re-imagining Archival Display: Creating User-friendly Finding Aids," *Journal of Archival Organization* 9, no. 1 (2011): 4–31.

[lv] However compellingly it is presented as being exemplary of the state of the art, OCLC's ArchiveGrid, for instance, is essentially a union catalog of digital records—both single-level catalog records and multi-level finding aids—describing the overwhelmingly analog contents of primarily North American archival repositories. See http://archivegrid.org/.

[lvi] See Anne J. Gilliland, *Conceptualizing Twenty-first-century Archives* (Chicago: Society of American Archivists, in press).

[lvii] See http://archivegrid.org/ for ArchiveGrid; http://www.archivesportaleurope.net/ for Archives Portal Europe; http://worldcat.org/ for WorldCat; http://www.oac.cdlib.org/ for OAC; and http://archiveshub.ac.uk/ for Archives Hub.

[lviii] See http://www.europeana.eu/.

[lix] See Daniel V. Pitti, "National Archival Authorities Infrastructure," accessed April 28, 2013, http://ecommons.cornell.edu/bitstream/1813/28718/7/Pitti_SNAC-NAAC_Cornell.pdf.

[lx] See http://socialarchive.iath.virginia.edu/NAAC_index.html for NAAC; see http://socialarchive.iath.virginia.edu/ for SNAC.

[lxi] See http://viaf.org/ for VIAF; see, e.g., http://lodlam.net/ for more on LOD; see, e.g., http://semanticweb.org/ for more on the Semantic Web.

[lxii] See, e.g., David A. Bearman, "Automated Access to Archival Information: Assessing Systems," *American Archivist* 42, no. 2 (1979): 179–190.

[lxiii] Lytle, "Subject Retrieval in Archives"; "Intellectual Access to Archives: 1. Provenance and Content Indexing Methods of Subject Retrieval," *American Archivist* 43, no. 1 (1980): 64–75; and "Intellectual Access to Archives: II. Report of an Experiment Comparing Provenance and Content Indexing Methods of Subject Retrieval," *American Archivist* 43, no. 2 (1980): 191–207. See also Bearman and Lytle, "The Power of the Principle of Provenance."

[lxiv] See, e.g., http://www.w3.org/XML/ for more on XML.

[lxv] See http://www.w3.org/TR/xpath/ for more on XPath; http://www.w3.org/TR/xquery/ for more on XQuery.

[lxvi] See, e.g., Mounia Lalmas, *XML Retrieval* (San Rafael, CA: Morgan & Claypool, 2009).

[lxvii] See http://www.loc.gov/ead/ for more on EAD.

[lxviii] See http://staff.science.uva.nl/~kamps/readme/ for more on the README project; see Junte Zhang, "System Evaluation of Archival Description and Access" (PhD diss., University of Amsterdam, 2011), http://www.illc.uva.nl/Research/Dissertations/DS-2011-04.text.pdf, for the dissertation.

[lxix] Zhang, "System Evaluation," 73.

[lxx] Bertram Ludaescher, Richard Marciano, and Reagan Moore, "Towards Self-validating Knowledge-based Archives," in *11th Workshop on Research Issues in Data Engineering* (Heidelberg, Germany: IEEE Computer Society, 2001), http://www.sdsc.edu/~ludaesch/Paper/ride01.html; Reagan Moore, Chaitan Baru, Arcot Rajasekar, Bertram Ludaescher, Richard Marciano, Michael Wan, Wayne Schroeder, and Amarnath Gupta, "Collection-based Persistent Digital Archives: Part 1," *D-Lib Magazine* 6, no. 3 (2000), http://www.dlib.org/dlib/march00/moore/03moore-pt1.html; Reagan Moore, Chaitan Baru, Arcot Rajasekar, Bertram Ludaescher, Richard Marciano, Michael Wan, Wayne Schroeder, and Amarnath Gupta. "Collection-based Persistent Digital Archives: Part 2," *D-Lib Magazine* 6, no. 4 (2000), http://www.dlib.org/dlib/april00/moore/04moore-pt2.html.

[lxxi] See https://inex.mmci.uni-saarland.de/ for more on INEX. See also, e.g., Mounia Lalmas and Anastasios Tombros, "Evaluating XML Retrieval Effectiveness at INEX," *ACM SIGIR Forum* 41, no. 1 (2007): 40–57.