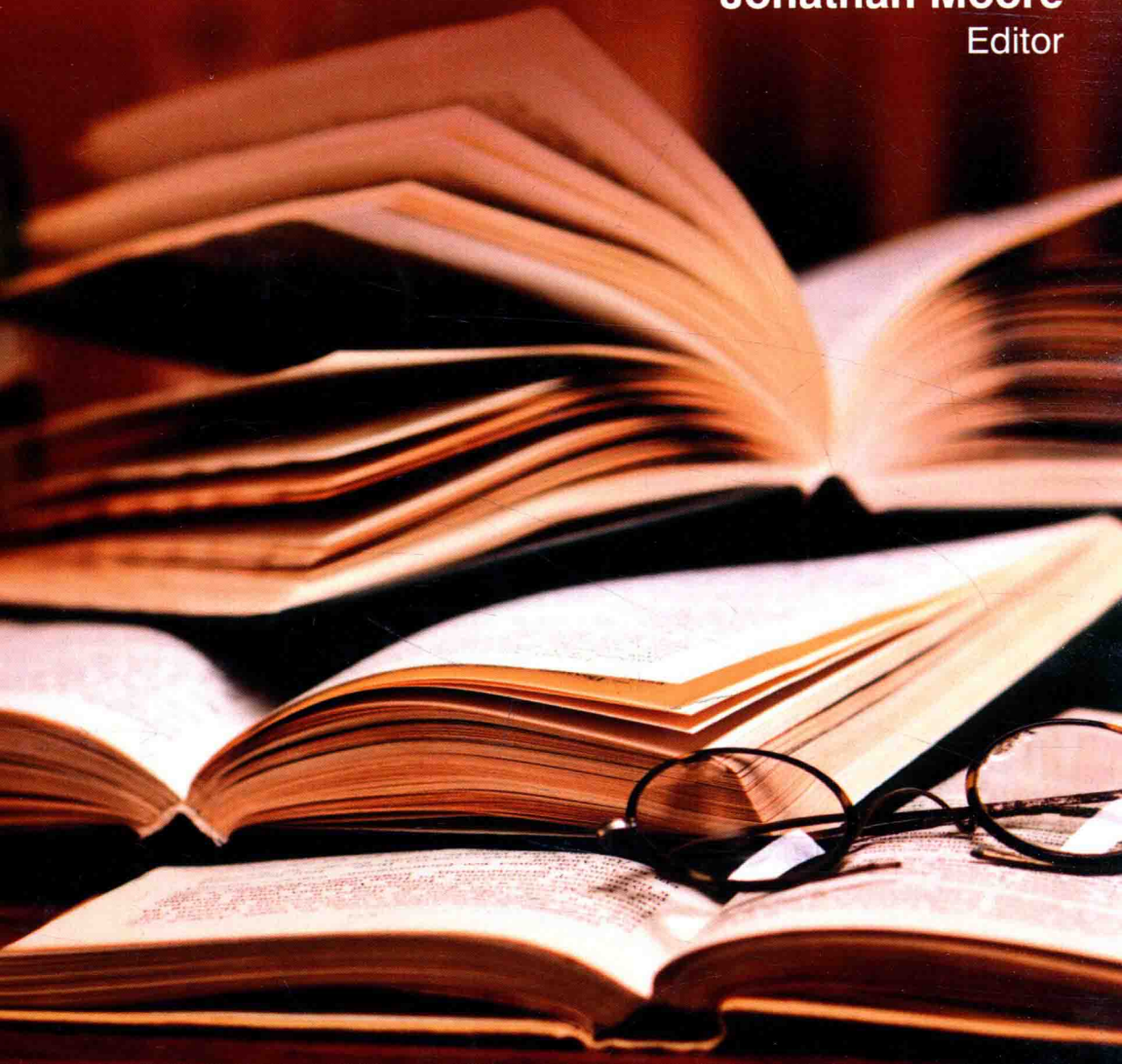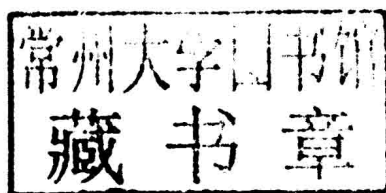# Research Methodologies in Library and Information Science

Jonathan Moore

Editor

Research Methodologies in Library and Information Science

Editor

Jonathan Moore

KORO PRESS

KOROS PRESS LIMITED

London, UK

Research Methodologies in Library and Information Science

© 2014

Research Methodologies in Library and Information Science

Preface

Library and information science (LIS) is a merging of the two fields library science and information science. The phrase "library and information science" is associated with schools of library and information science, which generally developed from professional training programs to university institutions during the second half of the 20th century. In the last part of 1960s schools of librarianship began to add the term "information science" to their names. The first school to do this was at the University of Pittsburgh in 1964. More schools followed during the 1970s and 1980s, and by the 1990s almost all library schools in the USA had added information science to their names. The trend was more for the adoption of information technology rather than the concept of a science. A similar development has taken place in large parts of the world.

Knowledge and experience of research is a fundamental part of what makes an information professional, and the possession of research skills is an essential tool for achieving success in the information environment. With the increasing need for accountability in enhancing service delivery and provision in the information workplace, research can be used to carry out benchmarking exercises, service planning and strategic development; to map user behaviour, plan operational changes, demonstrate value to stakeholders, and much more. Because LIS is an interdisciplinary field, researchers use multilple methods, depending on the nature of the question, the data sources, the setting, the timeframe, and more. It should be considered that information science grew out of documentation science and therefore has a tradition for considering scientific and scholarly communication, bibliographic databases, subject knowledge and terminology etc. Library science, on the other hand has mostly concentrated on libraries and their internal processes and best practices. It is also relevant to

consider that information science used to be done by scientists, while librarianship has been split between public libraries and scholarly research libraries.

This unique new handbook is the first to focus entirely on the research needs of the information and communications community. It provides a reference guide to the research process in the fields of information studies, communications, records management, knowledge management and related disciplines, and is designed to be of continuing use throughout the student and professional lives of its readers.

—Editor

Contents

Chapter 1

Open Architecture for the Confederated Digital Library

Introduction and Background

The emergence of the networked information system environment has allowed us to envision digital library systems that transcend the limits of individual collections to embrace collections and services that are independent of both location and format. An early effort to create a virtual collection based on distributed collections with a core set of services is the Networked Computer Science Technical Reference Library (NCSTRL).

NCSTRL is a confederation of over 100 institutions with the goal of providing a federated library of computer science material, *i.e.*, a seamless federation of collections and associated library services accessible to the broad community. NCSTRL encompasses three broad areas of activity:

1. The building of the NCSTRL federation of digital libraries and associated users,

2. The development of an open architecture as a technical basis for accomplishing the federation, which can also serve as a proposed architecture for more general federated digital libraries, and

3. The development and demonstration of a "reference implementation" of the proposed protocols/services: an instantiation of the architecture that both makes credible the approach and provides a software suite that federation members can use.

As an initial technical underpinning, NCSTRL used a combination of technologies from two prior projects: the DARPA-sponsored Computer Science Technical Report (CS-TR) project and the NSF-sponsored Wide Area Technical Report Service (WATERS) project. From CS-TR came architectural concepts, supporting middleware tools (*e.g.*, the handle system) and the Dienst system for disseminating, searching, and accessing material.

From WATERS came the tools and techniques for exploiting existing FTP archives of material and making them accessible in a uniform manner. Participating institutions have the option of either using the server software (prototype provided by Dienst), called NCSTRL-standard, or making their material available through the FTP approach of WATERS, called NCSTRL-lite. As NCSTRL evolves, additional technical components are being integrated, *e.g.*, the STARTS resource discovery protocol developed at Stanford University. In their recent D-Lib Magazine article, "Defining Collections in Distributed Digital Libraries", Carl Lagoze and David Fielding describe some of the current research at Cornell University that is extending the underlying architectural concepts.

The current status of NCSTRL may be found through the documents on the NCSTRL web site. Briefly, more than 100 institutions, mainly universities, are participating in NCSTRL by providing their material online either through an NCSTRL-standard server or through NCSTRL-lite access to their FTP archive. In addition, other forms of institutions (*e.g.*, the e-print server from Los Alamos National Laboratory (LANL) and *D-Lib Magazine*) are members of the federation, thereby adding material of types different from the original technical reports. Thus, a broad variety of material is accessible through the NCSTRL services including (but not limited to) technical reports, preliminary papers, and magazine and journal articles.

A key element of the NCSTRL activity is the development and demonstration of an *open architecture* approach to the *confederated digital library*. By open architecture, we mean that the functionality of the digital library is partitioned into a set of well-defined services or functions, each with a well-defined protocol specifying the interface to that service. Some of these services are intended to support users directly; some are intended for access by machines. Thus, organizations are free to develop or use different designs and implementations of the services, as long as their interfaces are consistent with the agreed upon protocol. This allows, for example, material to be served through a variety of repositories, such as the LANL e-print server, the FTP

servers at many institutions, and the Dienst servers at many members of the NCSTRL federation.

Furthermore, the architecture supports extending the capabilities of the system through specification of additional services. For example, extensions to support pricing of accessed objects are being explored. We note that achieving interoperability among digital libraries requires, in addition to conformance to an open architecture, agreement on items such as formats, data types, and metadata conventions. Some of these agreements will be embedded in the definitions of the protocol interfaces, but some won't. This extensibility is particularly important as the community struggles with the issues in achieving semantic interoperability. By confederated digital library, we mean that the organizations collaborating in the federation are autonomous and free to aggregate, at various levels of abstraction, whatever technologies they deem appropriate to satisfy their "customer" requirements, as long as they provide well-defined interfaces to their services that are consistent with the overall architecture. The confederated digital library therefore presents a seamless collection of digital library capabilities to the user, achieved through a federation of individual organization's digital library systems.

The open architecture supports the federation concept in two ways. First, it provides a modular construct for each organization to create its own digital library system, through selection of the best technology products to serve the needs of the local users. These products will interoperate to create the organization's digital library in a manner that is both sensitive to local conditions and needs as well as consistent with the overall federation. Second, the architecture provides a means for defining the interfaces between the digital libraries of the federation. This is achieved through specification of interfaces to the services provided by each of the separate libraries. The purpose of this brief document is to outline the overall approach being taken to the required open architecture for NCSTRL. The next section discusses some of the underlying assumptions that drove the NCSTRL architectural approach. Following that, an overview is provided of the approach being pursued in NCSTRL. Details on the various aspects of the architecture are left to the documents referred to in the discussion.

Underlying Assumptions

The notion of the federated digital library is the fundamental driver of the architectural approach being explored in NCSTRL. Each

of the organizations participating in the federation has a set of users that it needs to serve and/or a collection of material that it wants to make available. Currently, the typical institution in the federation is a university department or other computer science research organization that wants to assist in making the technical reports and other research material generated by their researchers available to the broader community in a way that maximizes the ease of dissemination, search, access, and retrieval. A core motivation of NCSTRL is to improve early and detailed communication of research results across the community. Thus much of the early focus of NCSTRL has been on preprints and other material that is worth disseminating but would not normally be published in a peer-reviewed journal or similar publication—material sometimes referred to as "gray literature". The material to be dealt with by NCSTRL, therefore, has great diversity, including software, documentation, technical reports, and white papers as well as more traditional journal-published material. A major advantage to the NCSTRL user is the ability to deal with (*e.g.*, search) this wide variety of material through a single interface.

NCSTRL represents a specific user domain of interest. However, as explained above, one of our goals is to develop an architecture that has broad applicability to federated digital libraries. We are therefore using a broad definition of digital library in the development of the architectural approach. The D-Lib Working Group on Digital Library Metrics in its description of the scope of a digital library, provides a relevant view of the intended scope of applicability of our work: "The Digital Library is the collection of services and the collection of information objects that support users in dealing with information objects available directly or indirectly via electronic/digital means." Each digital library in a federation such as NCSTRL faces the issue of designing its local system to respond to two drivers. One is to assure that local users can gain access to the material available in a manner most responsive to their needs.

The other is to make local collections of material and services easily and effectively accessible to the broad community, subject to policy and resource constraints. Thus, each institution may wish to design its own digital library system (aggregation of technology products providing the requisite collection of services), but at the same time, wishes to be "interoperable" with the broader community federation. The challenge in designing the open architecture, then, lies in providing means for both local autonomy as well as interoperability and composability of services.

Aspects of the Architecture

These drivers, and particularly the distinction between interoperability and composability, has led us to three aspects of the open architecture. The first is the common underlying infrastructure that supports the creation of multiple and extensible services. The second is the service decomposition—the partitioning of the digital library functionality into a set of well-defined services, each with a well-defined protocol specifying the interface to that service. The third is the mechanisms for interoperability among library systems that may not share the same service decomposition.

Supporting Infrastructure

Creating an open architecture for the federated digital library requires a certain degree of common supporting elements. In the Internet, the Internet Protocol (IP) provides a common addressing mechanism and basic packet format, thereby allowing all systems to move data among them. Digital libraries have an analogous requirement for being able to name digital information objects and resolve those names into addresses for retrieval of the objects.

Robert Kahn and Robert Wilensky, in their report "A Framework for Distributed Digital Object Services", define the basic approach to the supporting infrastructure that has been adopted in NCSTRL. The starting point is the notion of a *digital object*. "A digital object is a data structure whose principal components are digital material, or data, plus a unique identifier for this material ..." The structured data consists of other digital objects as well as elements which are not digital objects. The unique identifier in our system is called a *handle*, and the naming service is the *Handle System*.

The Handle System provides the means for associating identifiers (or names) with digital objects (the basic information elements of a digital library), associating addresses for those objects (such as Uniform Resource Locators, URLs), resolving queries from other parts of the system as to the address associated with the named object, and managing the overall system including organizational management (recognizing the autonomy of members of the federation in assigning names, for example) and management of name evolution (as objects change, for example). The final part of the core infrastructure (along with the notion of digital objects and a naming system) is a common *repository access protocol* (RAP). The RAP is to be supported by all repositories in the system, and defines the core set of interactions with that repository, such as storing or retrieving a digital object. RAP is

not an implementation blueprint, but rather only an interface description that is technology independent. The repository itself may be considered as a digital object containing other digital objects.

All three elements of the core infrastructure (digital object definition, naming service, and repository access) deal with digital objects as structured data and do not address the content of the objects. The core infrastructure deals with issues associated with the deposit, storage, access to and dissemination of the objects, thereby driving the need for the naming/addressing functions as well as definitions of aspects such as basic rights and permissions. It is left to the library services described in the next section to deal with aspects requiring knowledge of content, such as search. Meta-objects are digital objects that reference other digital objects for purposes of organizing and aggregating groups of digital objects.

One of the core tasks in the NCSTRL effort is to take the basic framework and approach to the supporting infrastructure described above and refine it into solid specifications suitable as a basis for an open architecture.

Services and Interfaces

Building upon the core supporting infrastructure, a set of digital library services are defined. The purpose of these services is to support users in dealing with the variety of content of the library—storing, retrieving, searching and aggregating information as required. Carl Lagoze and Sandra Payette, in their report "An Infrastructure for Open-Architecture Digital Libraries", describes the various services currently planned to support NCSTRL. The following figure and service descriptions are based on their definitions.

- The *repository service* provides the mechanisms for the deposit, storage, and access to digital objects. A digital object is considered *contained* within a repository if the handle of that object resolves to the respective repository (and, thus, access to the object is only available via a service request to that repository). The repository service provides more than simple deposit and access to objects, though, and can provide sophisticated management, aggregation, and marshalling of the information stored in the repository. As part of NCSTRL, Cornell University is exploring advanced repository architectures and services through its FEDORA effort.

- The *index service* provides the mechanisms for the discovery of digital objects via query. Individual index servers index

actual or surrogate information on sets of digital objects (which may be distributed across multiple repository servers). Queries submitted to these index servers return result sets that contain the handles of digital objects that match the query. The index service also provides metadata about the content of its indexed information and the capabilities of its query mechanisms. This metadata is used by other services, such as the collection service described below. The Stanford STARTS protocol design provides a basis for the design of such an index service.

- The *collection service* provides the mechanism for the aggregation of sets of digital objects into meaningful (from some community's perspective) collections. A collection server creates collections by, for example, scanning a set of index services, reading their metadata and applying its *collection definition criteria* to define which objects indexed by those index servers are elements of its defined collections. There is no fixed notion of collection definition criteria. One example of a collection definition criterion is *subject,* which may be determined by reading a controlled vocabulary metadata field or derived via some natural language analysis. The elements of a collection defined by a collection server may be indexed by any number of index servers and located in any number of repository servers.

- A *user interface gateway* provides a human-centred entry point to the functionality of the federated digital library. Each user interface gateway uses the information provided by one or more collection servers to permit searching for and access to objects within those collections. User interface gateways also use information provided by collection servers and index servers to make query routing decisions based on factors such as content, cost, performance, and the like. Thus, the gateway provides an easy mechanism for users (through the browser of their choice) to gain access to the variety of NCSTRL services in a consistent manner.

Underlying the services and their interactions are the common naming service provided by the handle system and the use of a common repository access protocol. Thus, any of the services would be able to determine the location(s) of an object through resolution via the handle system. Any of the services (*e.g.*, index services) would be able to access any repository in a standard manner through use of the repository access protocol. This architecture then permits

exploration and development of new enhanced versions of these (and other services) based on the underlying common framework coupled with a shared understanding of the basic functions provided by each of the services and how they interact. NCSTRL is exploring some of these alternatives (*e.g.*, FEDORA and STARTS).

Interoperability among Federated Libraries

An organization operating a digital library that is a member of the federation in essence has integrated a number of the services described above in support of its users. If this services integration is done in a manner consistent with the service architecture described above, and the various services conform to the interface specifications, then interoperability at a basic syntactic level is relatively straightforward to achieve. It can be done at multiple levels. For example, the repository service might be interoperable with index and collection services of other federation members through the well-defined interface. Or a collection service of one library may be accessible from the user interface of another.

However, it must be assumed that many digital libraries will have internal architectures that do not follow the NCSTRL approach. They may be based on a different service decomposition, or use different protocol specifications for internal communication among individual services. Thus, NCSTRL must also deal with interoperability on a more coarse scale. One approach being explored by Kurt Maly and the ODU Digital Library Group is based on a data driven architecture that allows integration of heterogeneous systems, and thus is insulated from changes in individual systems. A Digital Library Definition Language (DLDL) is being developed that is based on the Extensible Markup Language (XML). The DLDL will be capable of describing APIs for a wide variety of digital libraries. The richness of mark-up tags in a DLDL will be determined by the user needs or expectations from a federated digital library.

This approach makes it possible for a library to implement its own policies and features as well as to change them as long as it is able to describe these changes in the XML-based language. In particular, it does not require any existing library to change its architecture but only to describe it. The data-centred architecture being proposed consists of three major components: (i) a collection of heterogeneous digital libraries along with their descriptions in the DLDL, (ii) a registration service and the master XML merger agent, and (iii) a Federated Digital Library (FDL)—a Java based application that

facilitates the integration of different digital libraries and gives the user an impression of a single digital library.

The registration service allows a digital library to become part of a federated digital library by submitting its description in DLDL to the registration server. The DLDL description contains metadata for the digital library including its contents, and methods to interact with the digital library. For example, the type of the digital library, URLs for invoking various services of the digital library, and lists of associated members will be described using the DLDL. A digital library will be allowed to register, no matter how different it is from the others, as long as it uses the DLDL to describe its structure, methods, and behaviour, although it should be clear that in actuality there will be some management process for approving the registration.

Summary

This document has given a high level view of the approach being taken to achieve the open architecture required for the federated digital library of NCSTRL. We believe NCSTRL is an ideal testbed for such a federated library approach, and the overall approach has already been demonstrated to be feasible and useful. Our intent is to continue development and evolution of this architecture and the NCSTRL system. At the same time, though, we believe that this architecture is useful for environments other than NCSTRL and would welcome partnerships in exploring its applicability across the community as an open architecture approach to digital libraries and more generally to information management. Those interested in having such a dialogue are encouraged to contact the author.

Evaluating Search Engine Models for Scholarly Purposes

The Internet allows for the efficient dissemination of texts, thereby creating a rich hypertextual environment that is potentially conducive to stimulating the free exchange of ideas in a manner worthy of the modern scholar. However, the fact that any user whatsoever may disseminate texts in this manner presents two distinct problems. First, finding relevant resources on the Internet may take a fair amount of time and, second, once resources are found, determining their reliability is often difficult if the user is not already an expert in the field of the resource under consideration. These problems—efficiency in searching and academic quality-control—are surmountable with existing technology, and many laboratories around the world are working hard to shape this technology into a variety of academic information retrieval services.