STORIES

D-Lib Magazine
July/August 1999

Volume 5 Number 7/8

ISSN 1082-9873

Reference Linking for Journal Articles

Priscilla Caplan
University of Chicago
p-caplan@uchicago.edu

William Y. Arms
Cornell University
wya@cs.cornell.edu

Abstract

During the past year, great progress has been made in the field of reference linking, particularly in the important area of links to journal articles. This paper summarizes the current state-of-the-art, describes a general model for static linking, compares several current implementations against the model, and discusses some of the required future work. Particular emphasis is given to the minimal set of metadata needed for reference linking and to selective resolution of identifiers, methods by which a client can specify which of several copies of an item is accessed.

Introduction

Reference linking is the general term for links from one information object to another. The links may appear in a wide variety of contexts, including published citations to scientific works, references from a catalog or bibliography, and informal references transmitted by email or verbally. In recent years, extensive development has been carried out on reference linking between journal articles, and recently work has gone beyond journals. One of the first projects to examine reference linking systematically was the Open Journals Project [Hitchcock 1998].

Recently, several systems have been developed for reference links from online journal articles to other journal articles. The most complete, within its limited domain, is provided by the NASA Astrophysics Data System [ADS]. Another leading example is the National Library of Medicine's PubMed/PubRef [PubMed] system, which is used by HighWire Press and others. An excellent commercial example is ISI's Web of Science [Atkins 1999]. The International DOI Foundation (IDF) is leading another effort, using Digital Object Identifiers (DOI), a form of Uniform Resource Name [Paskin 1999].

In February 1999, the National Information Standards Organization (NISO), the Digital Library Federation (DLF), the National Federation of Abstracting and Information Services (NFAIS), and the Society for Scholarly Publishing (SSP) sponsored a one day invitational workshop to discuss issues surrounding reference linking, specifically linking from citations to electronic

journal literature. The report of the February linking workshop is available at [Needleman 1999]. The participants identified three major components for constructing systems to support reference linking: identifiers for the works; a mechanism for discovering the identifier of a work from a citation; and a mechanism for taking the reader from an identifier to a particular item. A small working group was assembled to review, refine, and elaborate on the work of the first workshop. Their report [Caplan 1999a] was the basis of a follow-up workshop in June [Caplan 1999b]. This paper is an elaboration of that report. It places the results of the workshops within a broader discussion of the current state of reference linking.

The generic statement of the reference linking problem is, "Given the information in a standard citation, how does one get to the thing to which the citation refers?" The major focus of the workshops, however, was citations to journal articles. Thus, the problem statement for the meeting of the working group was, "Given the information in a citation to a journal article, how does a user get from the citation to an appropriate copy of the article?" The working group was explicitly asked to consider the situation where there are several copies of an item and the user may have a preference for which item copy is supplied. The group coined the term "selective resolution" for this situation.

The hyperlinks of the web, using URLs, often perform as surrogates for reference links. Hyperlinks can be used to represent citations, to structure information, or for a myriad of related purposes, but they suffer from several disadvantages when used as reference links. A URL identifies a single instance of a work, not the work itself. Since URLs reference a specific location, they are vulnerable to changes or poor management of the system at that location. Hence, research on reference linking is allied to the development of systems of persistent identifiers.

Throughout the study, the emphasis has been pragmatic. What is needed to get started? Are there simplifications that can be made in the short term, knowing that they will need to be addressed later? However, reference linking goes much further than citations to journal articles, and the simplifications that are being used to get started must always be considered in the long-term context. (See the discussion of dynamic linking below.)

Creations

The first stage in reference linking is to understand to what a reference refers. The framework from the IFLA report, "Functional Requirements for Bibliographic Records", provides a vocabulary for distinguishing between related aspects of an intellectual entity [IFLA 1998]. In the IFLA model, a "work" is an abstract conception of some creator. Works are realized through "expressions", which are fixed spatial/temporal representations of works, such as a performance of a play or a symphony. Expressions in turn are embodied in "manifestations", physical representations such as printed books or recorded CDs, which may or may not be mass-produced. A specific, single manifestation is an "item", also called a "copy".

The European INDECS project has done a careful analysis of these distinctions and proposes a categorization that, while somewhat different from the IFLA model, is mainly compatible with it [INDECS]. Supplementing the

IFLA and INDECS terminology, the International DOI Foundation (IDF) has contributed "creation" as a useful generic term encompassing the work and all of its expressions, manifestations and items.

The distinction between expression and manifestation is useful for works that are performed but usually can be ignored for works that have a single expression, like most journal articles. Journal articles represent three types of creations: the work, or creative output of the author(s); the manifestations, or instantiations of the work in print and/or electronic form; and the items, or specific copies of a manifestation. An article, for example, could have been published in a print and an electronic version. These would be separate manifestations, each of which might have multiple items (perhaps several hundred copies for the print run, and mirrored online and archival copies of the electronic version).

Citations and creations

The author of a citation sometimes refers to a work, sometimes to a specific expression or manifestation, and sometimes to an individual copy. Often a citation will refer to a specific manifestation only because the citer, working from his own copy of the article, is unaware of other manifestations that would do as well.

In some cases, however, an author will cite a particular manifestation deliberately. The *British Medical Journal* provides an example of a publication where manifestation is significant. Articles are published in three manifestations: print, PDF, and HTML. For some articles, the print and PDF are abridged versions of the full HTML article, which may be longer, and may contain additional figures and references. However the official citation given by the publisher refers to the print/PDF manifestations, including the pagination, which is not relevant to the HTML.

Consideration of the *British Medical Journal* leads to the question of under what circumstances the different versions should be considered different works, as the intellectual content varies. The distinctions between work, expression, and manifestation are a matter for judgment. The IFLA model is analytic while publishers are declarative, in essence defining different manifestations as distinct or equivalent by declaring that they consider them so. This example illustrates that the IFLA model must be seen as a general framework rather than a precise definition or specification.

In the absence of a clear indication of the author's intentions, it can usually be assumed that a citation refers to the work, as both the citer and the reader can be expected to be primarily interested in the intellectual content. (This is true even though when a citation uses a URL, the author is usually constrained to refer to the location of a specific copy.) Most current implementation projects focus on citations to works, and hence on the association of identifiers with works, while recognizing that occasionally there will be a need to distinguish different manifestations. This is the approach taken by the Astrophysics Data Center, ISI, and PubMed. One of the central aims of INDECS is to be explicit in distinguishing between the underlying work, its various expressions, and its manifestations. The IDF is a member of INDECS and is bravely attempting to be explicit about the distinctions, but has accepted that its initial services can

refer generally to "articles". Currently, this cautious pragmatism seems an acceptable simplification.

A general model for reference linking of journal articles

Although they differ greatly in details, most current systems fit within the framework shown in Figure 1. (A notable exception is SFX, which is mentioned briefly below.)

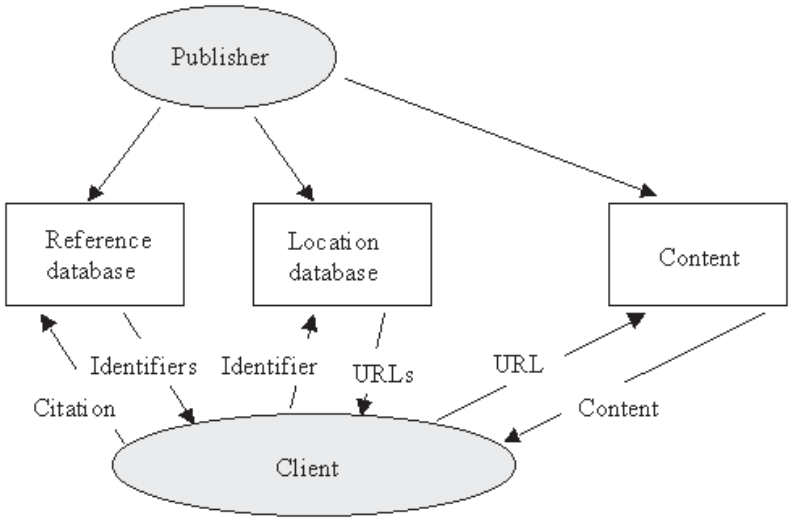

Figure 1. Reference linking

Each work has a unique identifier and one or more copies, each with its own URL. The provider of the information, who is usually the publisher, supplies metadata about each work. This is stored in databases as shown in the middle row of Figure 1. Clients access the databases through the interactions shown in the bottom part of the figure. The figure shows two databases: a reference database and a location database.

Reference database

> For each work, the reference database contains metadata that, at a minimum, corresponds to the information in a conventional citation. A client that wishes to find the content associated with a reference sends a query to the reference database. This database returns a list of identifiers for works that match the query.

Location database

> Typically each cited work will be stored at several locations. A client sends an identifier to the location database, which returns one or more URLs. The client selects the URL to retrieve the object. This is known as "resolution" of the identifier.

This process has many complications. There will be considerable variation in citations; some will be formally published as references within scholarly journal articles; some will be formulated as part of more casual

communications such as course reading lists and informal bibliographies. In some cases a citation may contain the identifier of the article explicitly, in which case the reference database lookup is not needed; in other cases an identifier will have to be obtained by using the bibliographic data elements given in the citation. There may be several works in the reference database that match the query; the client must select a work either by human intervention or by algorithm. When there are several URLs to different copies of the work, the system is faced with selective resolution: the client may wish to select a specific version based on variations of content, different licensing arrangements, or network performance.

Current implementations present several variations on this model. The Astrophysics Data Service derives references algorithmically, bypassing the reference database lookup. PubMed and the Web of Science combine the citation and location databases. Currently, all location databases return a single URL, though this is changing. PubMed's LinkOut experiment permits users to provide URLs in addition to those provided by publishers. The Handle System, which resolves DOIs, has an unused service that is capable of returning several URLs or other resolutions of a DOI.

Identifiers

An important question is whether effective reference linking needs identifiers other than URLs. The need for persistent identifiers has been widely advocated in a broader context than the reference linking problem. (See, for example, [Sollins 1994].) Yet, it can be argued that the deployment of general purpose Uniform Resource Names (URNs) has been slow and that wonderful systems have been built on the web using nothing more than URLs.

While it might be possible to build a reference linking model that does not presume the existence of identifiers, this seems unwise. Use of identifiers improves the reference linking model in a number of ways. Identifiers associated with works provide the primary means of clustering multiple copies of those works. The existence of the identifier allows citation lookup and resolution steps to be performed by different software systems, and facilitates distributed resolution. It provides management benefits for those running reference lookup and resolution services. Above all, the identifier gives permanence to a reference beyond the life span of any particular computer system. Given the overwhelming practical benefits of the identifier, it seems best to treat identifiers as a necessary part of the general model, while acknowledging there may be special cases in which they can be omitted.

Perhaps the most compelling argument that identifiers are needed for reference linking is that all current systems find them necessary. For ISI the identifier is a private key. The Astrophysics Data System has its own BibCode, and PubMed uses a PubMed ID. Digital Object Identifiers (DOIs) are an implementation of a Uniform Resource Name; they are public identifiers intended to be used wherever the item needs to be identified. DOIs are managed and resolved through the CNRI Handle System [Handle]. BibCodes and PubMed IDs were not explicitly intended to be Uniform Resources Names, but can be considered as such. They satisfy the commonly accepted criteria of persistence and global uniqueness, while supported by openly-accessible resolution systems.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.