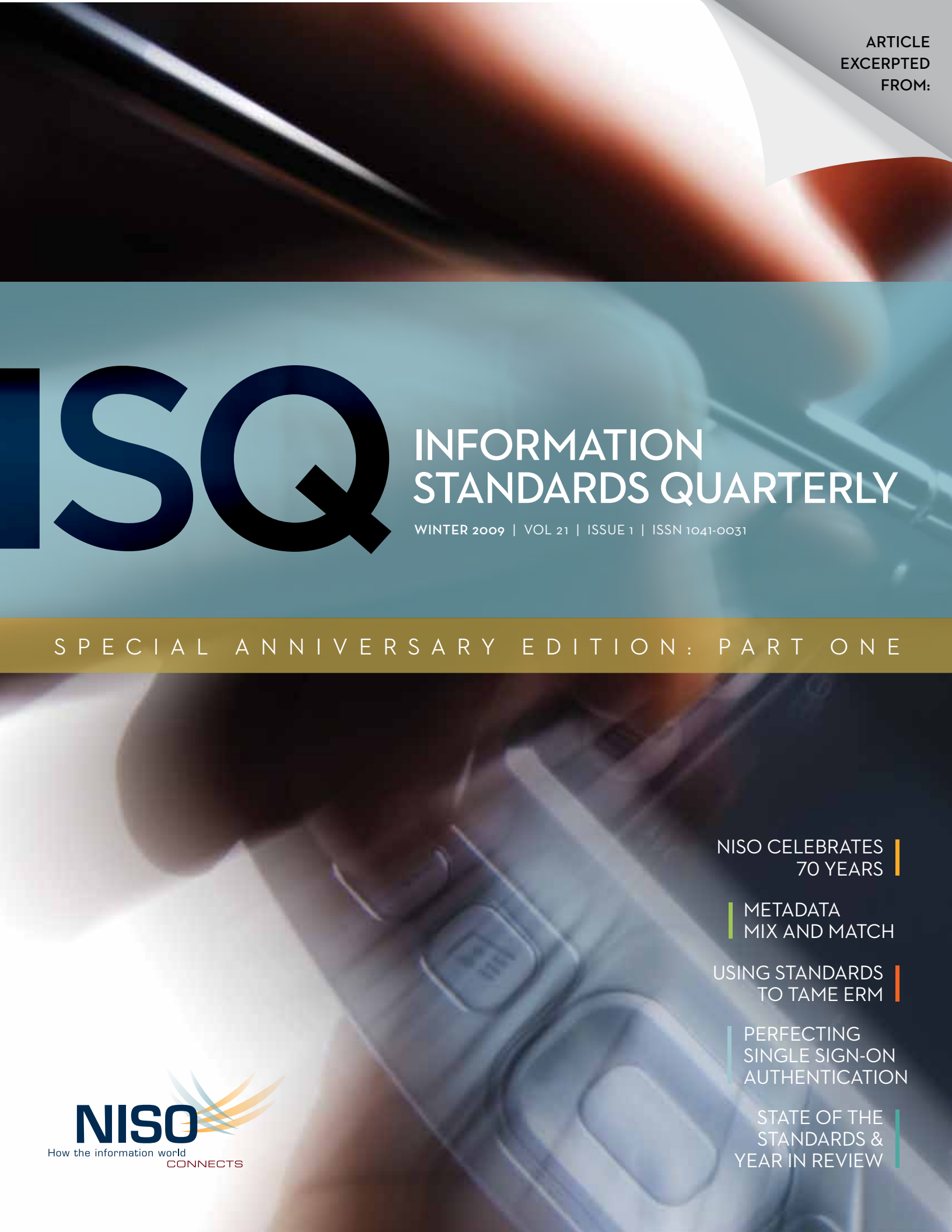


ISQ

INFORMATION STANDARDS QUARTERLY

WINTER 2009 | VOL 21 | ISSUE 1 | ISSN 1041-0031

SPECIAL ANNIVERSARY EDITION: PART ONE

NISO CELEBRATES 70 YEARS

METADATA MIX AND MATCH

USING STANDARDS TO TAME ERM

PERFECTING SINGLE SIGN-ON AUTHENTICATION

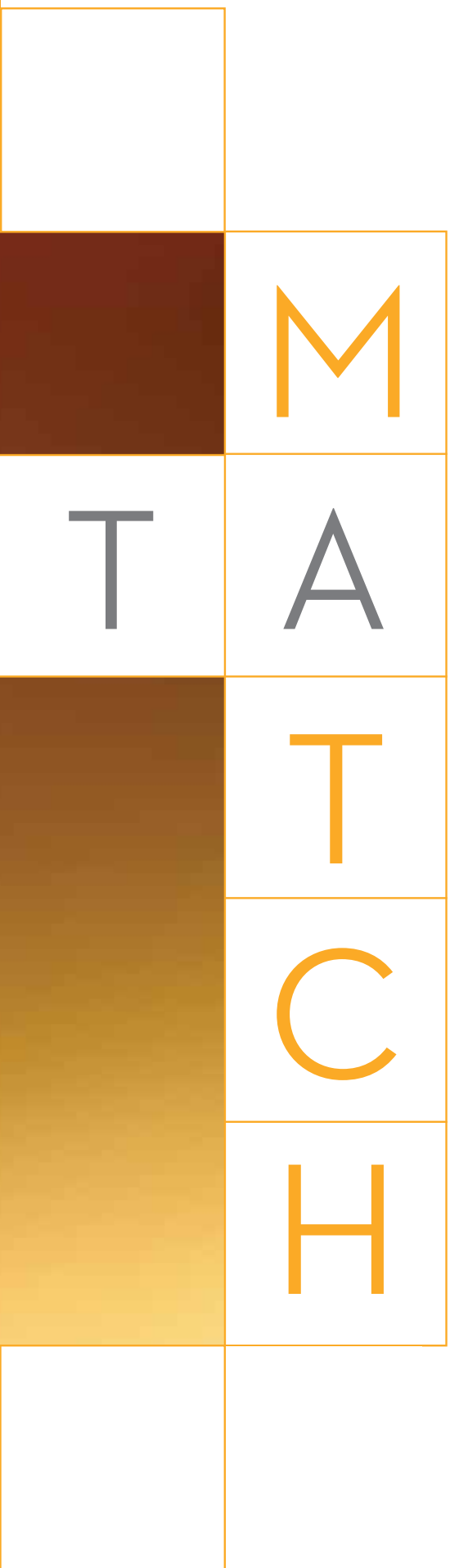
STATE OF THE STANDARDS & YEAR IN REVIEW

NISO
How the information world CONNECTS

AN METADATA IX

KAREN COYLE

It's not uncommon for those of us associated with libraries and library bibliographic data to think of bibliographic metadata as being specifically a record. It's also not uncommon for us to think of only one kind of record: the one we now call "MARC 21." In fact, our metadata standards generally define records as the unit of the standard, including the early NISO standard, ANSI Z39.2, *Information Interchange Format*, which defines the underlying format for the MARC 21 record.



Increasingly, however, I am finding that the record view doesn't match the complex bibliographic reality that we live in today. Work that I have been doing on the Open Library, a project by the Internet Archive, has helped me come to the conclusion that our future is about data, not records, and that our applications must be able to work with a mixture of data standards.

Beyond Library Bibliographic Data

I was asked to consult with the Internet Archive's Open Library project primarily to lend my expertise in bibliographic data. At the time that I stepped in, there was a database design and a database with some bibliographic data. Although I've never been a cataloger, I have spent decades working with library data in MARC format, and I therefore have some pre-conceived notions of what bibliographic data should look like. To my dismay, the Open Library data did not look anything like library bibliographic data. I learned, however, that there were some good reasons for this.

The first was that the Open Library was not limiting itself to library data. In fact, a great deal of the data in the database comes from other sources, including data obtained from Amazon.com, ONIX data from some individual publishers, and even some records that have been hand-keyed by Open Library users. The default user view of the bibliographic data is a combined display of elements from the various sources, yet it is also possible to drill back through the history of the bibliographic entry to see all of the data that has been submitted, including each change that has taken place. The bibliographic entry is not a fixed item but a growing organism whose evolution is visible.

Another reason the Open Library does not limit itself to the more rigorous library data style was that the Open Library allows editing of its data by the general public: people with no particular bibliographic training. It is obviously not possible to present concepts like "country of producing entity for archival films" or even "uniform title" to an untrained user base.

The Open Library programmers were not familiar with the standard library metadata record, and the standards were not compatible with the general suite of tools that the programmers commonly work with, such as HTML, CSS, and a host of XML-based tools. Although most of the team's communication is via e-mail or chat (the project's personnel are on three different continents), I could hear the virtual sighs as I explained the nature of the MARC record and of the MARC-8 character set. Fortunately, generous souls in the library community provide translation routines from MARC into XML and the Unicode standard character set.

Link Data, Not Records

The most compelling reason to deviate from the standard view posited by library bibliographic data, however, has to do with the concept of linked data. It is expected that data today will interact with a wide range of information resources. The Open Library uses an underlying data design that is commonly called a "triple store." In this design, data elements are simple key/value pairs that can be re-combined for a variety of uses. The individual units, such as "author = John Smith," are available to be used as needed in whatever context is appropriate. The emphasis is on the data, not on a particular record. Freed from a particular record structure, the data is also available to link out to similar data in other data stores. For example, any persons named in the Open Library database can be linked to entries in Wikipedia for that person or to a personal web page. It doesn't matter that each of these resources has a very different overall structure

CONTINUED »

and may share only that one data element in common. When you emphasize data, rather than records, the different information sources reveal themselves to be less different than you may have thought.

It's true that the data presented by Amazon and the publishers is oriented toward the immediate marketing needs of those organizations, while the library data takes a longer and broader view of the bibliographic universe. But semantically, the similarities outweigh the differences, particularly in the eyes of the users, who easily understand these two entries to represent the same book:

1. **Run for Your Life**
   James Patterson
   In Stock
   Little, Brown and Company
   February 2, 2009
   Hardcover

2. **Author: Patterson, James, 1947–**
   Title: Run for your life : a novel / James Patterson and Michael Ledwidge
   Imprint: New York : Little, Brown and Co., 2009
   1 copy on shelf

The first is from Amazon, the second from a library catalog. Each in its own input record format is very different, but the data itself is more alike than it is different.

You can take advantage of both the similarities and the differences when you can store the data apart from any particular record format. For example, your author data can take multiple forms, each one being an authoritative form of the author's name in a particular context:

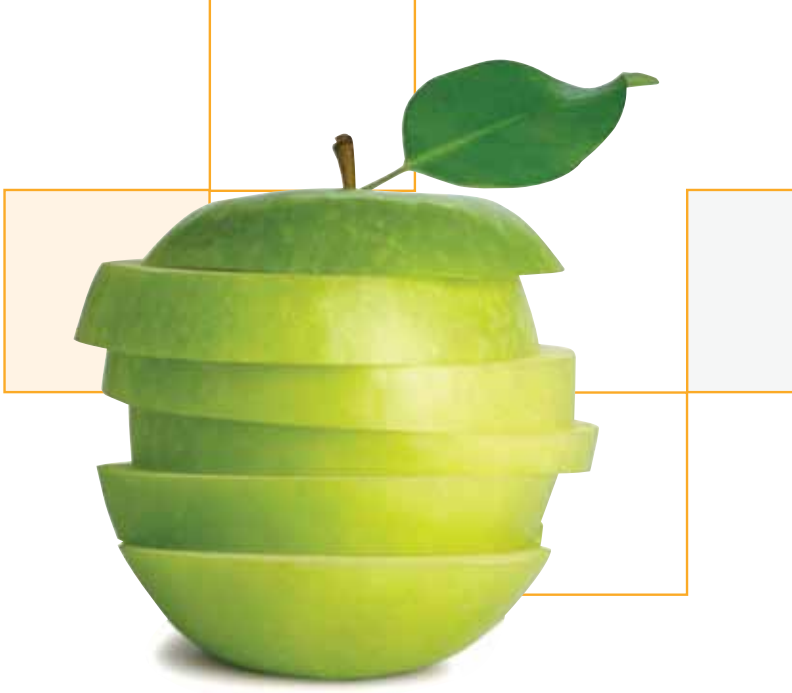
http://www.amazon.com/James-Patterson/e/B000APZGGS
↳ Displayed as: James Patterson

lccn:n78086409
↳ Displayed as: Patterson, James, 1947-

http://openlibrary.org/a/OL22258A
↳ Displayed as: James Patterson 1947-

http://en.wikipedia.org/wiki/James_Patterson
↳ Displayed as: James Patterson

Each of these is standard in its own environment, and each can be considered standard outside of that environment if it is identified clearly as to its source and has a unique identifier within that source. For these purposes, Uniform Resource Identifiers (URIs) are ideal, but other identifier formats can still be useful.



The "smart up" method allows you to merge and modify data using the best information you have.

Smart Up, Not Dumb Down

It's an unfortunate fact that many systems combine data from different sources using only the "dumb down" method, reducing the metadata to the few matching elements and resulting in the least rich metadata record possible. This results in a tremendous loss of data and an inferior user experience. The "smart up" method uses all or most of the data from the different sources, resulting in enhanced information. For example, the Open Library record is able to link to any number of information sources both from its pages for books and its pages for authors, in part because it can store linkable data from any source without having to be concerned about fitting that data into a particular record format. It also means that it can create a display that is richer than any one data source. The web pages for books combine subject headings from library data as well as the publisher's BISAC subjects. The web pages for authors can carry the biographical information that publishers include in their marketing data, yet can still be linked to name authorities records used by libraries to record the decisions about the author's identity.

The "smart up" method also allows you to merge and modify data using the best information you have. As we all know, matching the names of persons across systems is highly problematic. Although libraries put a great deal of effort into the identification of named persons and of corporate entities, the name forms that they choose to use as identifiers are not the ones used by any other community. Combining information from many sources allows you to make inferences based on the context of the data, so author names that are similar, though not identical, but share links to titles and publication information can be brought together

as likely matches. The more of this contextual data you have, the more sure you can be that your matches represent the same resource.

Metadata Dynamics

Once you accept that metadata does not have to represent a single source of data or a single defined record format, it becomes easier to see that metadata can be dynamic—that it can exist in multiple versions or in an assortment of views at the same moment in time. The Open Library uses the Wiki concept of change control, capturing each change to its content as an addressable web page.

Because of the mashed-up nature of the Open Library display, it is important to consider the original data sources as a continuing part of the information product. The design for the eXtensible Catalog (XC) is built around this same capability, facilitating both an incremental development of applications, but also potentially allowing the development of multiple applications from the same set of data. The days in which we discarded everything but the most recent version of a database record are over; versioning is in, which means keeping a history of all input and all changes to the bibliographic data. Ideally, it also means knowing where each data element originated, thereby retaining the ability to recreate a coherent, standards-based record when needed.

Mix and Match Metadata is the Future

It may seem that the Open Library is an anomalous project, and therefore not one that provides lessons we can apply elsewhere, but I see evidence that this type of project is in fact the new norm. Increasingly, we will be creating information services that accept and manipulate data that comes from multiple sources, each one based on different standards or no standards at all. We can plan for that eventuality, as evidenced by the XC project, but this means making a shift in our thinking about metadata. In particular, we need to move from an emphasis on records to an emphasis on data.

Much of what has been possible in the Open Library is because its main inputs—the library and the publisher data—themselves are heavily populated with standardized elements. It's clear that a data store can be open, dynamic and still adhere to standards, as long as the standards are applied to individual data elements. As we move more toward linked open data, it becomes vital that data elements adhere to standards so that they will be usable in a variety of contexts, or at least outside of the one context of the originating system. Those of us creating and using bibliographic data will need to develop a shared set or sets of element standards that are well-defined and web-ready. This means basing our data on data standards, not record standards. Examples of data standards are the Resource Description Framework (RDF) and Simple Knowledge Organization System (SKOS) of the World Wide Web Consortium, and the foundation standards of the Dublin Core Metadata Initiative, in particular the Abstract Model and the model for Application Profiles.

I would wager that we are seeing the end of the "pure" library cataloging record that contains only library-provided data. The future will be about data more than records, and the data will come from heterogeneous sources. This requires us to be more thorough in our data definitions, but also to design data knowing that it will have uses independent of a single, controlling record. This has important implications for how we engage in standards development from this point forward. We should no longer be defining data that is bound to a single record, but should be considering the broader context in which our applications and our data will interact. Not every data element will have a sibling in Wikipedia, but we should begin our standards work with the assumption that no data need is an island, and that no community has the only voice on any topic. | FE | doi: 10.3789/isqv21n1.200905

KAREN COYLE <www.kcoyle.net> is a librarian and a consultant in the area of digital libraries. She worked for over 20 years at the University of California in the California Digital Library as a developer specializing in metadata. Karen has served on library and information standards committees, including the MARBI committee advising on MARC standards, the NISO OpenURL committee, and currently the NISO Architecture Committee. She writes and speaks frequently on technical topics ranging from metadata development, technology management, system design, and on policy areas such as copyright and privacy.

RELEVANT LINKS

Dublin Core Metadata Initiative
www.dublincore.org
www.dublincore.org/documents/abstract-model/
www.dublincore.org/documents/singapore-framework/

The eXtensible Catalog (XC)
www.extensiblecatalog.org

MARC 21
www.loc.gov/marc

Open Library
www.openlibrary.org

Tim Berners-Lee on Linked Data Design
www.w3.org/DesignIssues/LinkedData.html

Wikipedia
en.wikipedia.org

World Wide Web Consortium
www.w3.org

w3.org/RDF
www.w3.org/2004/02/skos/