NISO RP-2005-03

NISO Metasearch Initiative

Search and Retrieval Citation Level Data Elements

A Recommended Practice of the National Information Standards Organization

Standards Committee BC / Task Group 3

Version 1.0
September 13, 2005

Published by the National Information Standards Organization
Bethesda, MD

Summary

The NISO Metasearch Initiative, Task Group 3/SubGroup 3, on Required Citation Metadata has discussed the issues around citation metadata and its relation to metasearch. Citation references have been devised in a paper world, assuming page numbers and enveloping journals and publishers. But searchers will use metasearch engines to search, find, and retrieve individual articles. There are a number of extant issues that must be addressed to allow smooth and seamless metasearching across multiple resources. The Google™ Scholar approach is to access the full-text content of all available journals and provide a heterogeneous data store. Unfortunately, for researchers, they need fine tuning of their search experience with relevant metadata so as not to be swamped by irrelevant references. Our proposed approach is simply to have a consistency in the format and content of citation metadata.

Issues

Inconsistent Citation Styles

The reference styles for citations tend to differ according to discipline. There are tens, if not hundreds of styles. As an example, one vendor has seventeen citation formats across twelve databases. The ISO and NISO standards are not in themselves a sufficient guide to all the variations. From the Dublin Core Metadata Initiative Citation Working Group, we get the following list of variations:

- The order of elements (especially elements such as initials)
- The mandatoriness of elements (e.g. many chemistry styles leave out the article title, but biology and medicine wouldn't)
- The punctuation between the elements
- Capitalization. E.g. of titles - some styles use "title case" (i.e. initial capitals for all main words) and some use "sentence case" (i.e. initial capitals for first word and proper nouns and adjectives only)
- Acceptable abbreviations (especially regarding journal title abbreviations, but also element indicators such as "chapter/chap/ch", "editor(s)/edited by/ed(s)", "edition/edn/ed"
- Character formatting (i.e. what goes in italic, bold, etc.)

Refer to http://epub.mimas.ac.uk/DC/citstyles.html for more discussion and a list of citation styles.

One of the reasons behind this plethora of styles is that data vendors purchase data from different publishers, each using potentially different styles.

Complex Technology Required

Due to the wide and varying citation formats returned by various vendors, metasearch engines must choose how to parse each citation. With “random” fields, even the parsed results are unreliable and inconsistent, oftentimes producing bad OpenURLs which can make it difficult for users to get to the full-text or article that was originally published.

Vendor Branding

Vendors and publishers desire to maintain their branding and identity in results sent to users, even after being massages by a metasearch engine. Either a vendor produces a proprietary OpenURL that will only point back to their own sources, or a vendor or publisher's reference is lost from the metadata. The vendor wants more exposure, renewed subscriptions, and possibly pay-per-view of full-text.

Mapping of Metadata

One issue that causes confusion and difficulty in de-duping records is the process in which multiple metadata items get placed into databases. A typical scenario goes as follows: the primary publisher creates a human readable citation field; the human readable citation field is dumped into a single database field; and the record in the database is sold to an aggregator. Since many different formats may be managed by one citation aggregator, it is difficult to tell which format they used for each citation. When the record is searched, it may be displayed as created by the publisher and not the authors.

Requirements

The requirements to enable effective and seamless metasearch across multiple databases and resource types are surprisingly simple. There are basically two audiences to the results of a metasearch: a metasearch engine, and the end-user. The combined minimum requirements end up being as follows:

Minimum metadata to allow a metasearch engine to compare results from multiple resources:

- Unambiguous metadata
- Enough to be able to Sort/Merge/Dedupe (OpenURL)
- Display (Brief/Full) – minimum for the user
- Produce OpenURL/Link
- Ranking: Need searched fields: Subject/Description/Abstract

To create a "Brief" Record, you need, at a minimum:

- Genre – what "type" of item is it?
- Creator – who created the original article?
- Title – how is this article referred to?
- ID – what ID(s), such as PII, SICI, DOI, etc., is this article known by?
- Context – what enveloping publication or proceeding, etc., is this article found in?

To create a "Full Display" Record, and to enable ranking and full-text analysis of the metadata, you need:

- Subject – for cataloged subject headings
- Description – some text describing what this item

Proposed Solution

A detailed table describing the minimum data elements needed for citation metadata follows this summary; an XML version of the table is available on the NISO Metasearch Initiative website (http://www.niso.org/committees/MS_initiative.html). This set is taken extensively from Dublin Core 0.1, qualified for citations from the citation working group, however, it adds the descriptive components needed for "Full Display" and text analysis done by metasearch engines.

A quick overview follows. As expected, it closely matches the Requirements listed above.

- "genre" element that describes WHAT kind of object we have
- an "authors" field, as in OpenURL
- "titles" field that has Journal Title and Article Title
- "dates" field that has the date of publication, and other chronological information if present
- "context" field that gives volume, issue, pages, etc.
- "citationID" for ISBN, ISSN, SICI, etc.
- "publisher" field, if available
- "fulltextURI" to point to the full-text, if available

For full display information, add the following. (If the information is requested by a metasearch server that is doing independent ranking of results, then this information is highly recommended to aid in the ranking of results.)

- "description" as in Dublin Core, for description or abstract
- "subject", as in Dublin Core, for subject headings
- "vendorData" — to include, in "free form" with a schema pointer, whatever else they want to add (This allows vendors to preserve branding.)

Links

For comparison and related links, here are other, similar standards, and a few discussions of interest:

Dublin Core Metadata Initiative Citation Working Group
http://dublincore.org/groups/citation/

Guidelines for Encoding Bibliographic Citation Information in Dublin Core Metadata
http://www.dublincore.org/documents/dc-citation-guidelines/

IMS Resource List Interoperability (RLI) Information Model, e-Learning metadata
http://www.imsglobal.org/rli/rliv1p0/imsrli_infov1p0.html

Digital Objects Requirements: Metadata, California Digital Library
http://www.cdlib.org:8081/inside/diglib/guidelines/mdreqs.html#mdguidelines

XBib - Bibliographies and Citations for XML
http://xbiblio.sourceforge.net/

RIS Format Specifications (EndNote)
http://www.refman.com/support/risformat_intro.asp

Identifier Encoding Schemes
http://epub.mimas.ac.uk/DC/citids.html

ANSI/NISO Z39.56-1996 (R2002), Serial Item and Contribution Identifier (SICI)
http://www.niso.org/standards/standard_detail.cfm?std_id=530

Marc Proposal No:2003-03, Definition of Data Elements for Article Level Description, Library of Congress
http://www.loc.gov/marc/marbi/2003/2003-03.html

Cameron, Robert D., Towards Universal Serial Item Names, School of Computer Science, Simon Fraser University, CMPT TR 97-16
http://www.cs.sfu.ca/pub/cs/TR/1997/CMPT97-16.html

Green, B., and Bide, M., Unique Identifiers: a brief introduction
http://www.bic.org.uk/uniquid.html

Registry for the OpenURL Framework - ANSI/NISO Z39.88-2004
http://www.openurl.info/registry/

Metasearch Search and Retrieval
Citation Level Data Elements

v 1.0

Element				Description
citation				The required root element "citation" contains child elements that are used to express properties of serial publications
	genre			Genre of the document. Legitimate values for the "genre" element are: (1) "journal": for a serial publication issued in successive parts (2) "issue": for one instance of the serial publication (3) "article": for a document published in a journal. (4) "conference": for a record of a conference that includes one or more conference papers and that is published as an issue of a journal or serial publication (5) "proceeding": for a single conference presentation published in a journal or serial publication (6) "preprint": for an individual paper or report published in paper or electronically prior to its publication in a journal or serial (7) "book" for monographs (8) "bookitem" for parts of a book, such as a chapter (9) "unknown": use when the genre is unknown.
	creator			The "creator" element contains child elements that are used to express authorship of an individual item in a publication. The "creator" element is not repeatable, it contains all authors, and allows for the indication of the position of the author in the publication's list of authors
		attr:rank		An integer indicating the position of the author in the publication's list of authors , e.g. "1" for first author, "2" for second author, etc.
		author		The person primarily responsible for creating the intellectual content of the resource
			aulast	The author's family name. This may be more than one word. In many citations, the author's family name is recorded first and is followed by a comma, i.e. Smith, Fred James is recorded as "aulast=smith"
			aufirst	The author's given name or names or initials. This data element may contain multiple words and punctuation, i.e. "Fred F", "Fred James"
			auinit	The author's first and middle initials.
			auinit1	The author's first initial.
			auinitm	The author's middle initial.
			ausuffix	The author's name suffix. Qualifiers on an author's name such as "Jr.", "III" are entered here. i.e. Smith, Fred Jr. is recorded as "ausuffix=jr"
		au		The author's full name, i.e. "Smith, Fred M", "Harry S. Truman"
		aucorp		Organization or corporation that is the author or creator of the book, i.e. "Mellon Foundation"
	title			The "titles" element contains child elements that are used to express the fully qualified title of an individual article in a serial publication. The "titles" element is not repeatable, it contains the journal or abbreviated journal title, article title
		atitle		Article title Either the Journal Title "jtitle" or abbreviated journal title, "stitle" must be supplied. However, the data vendor may supply both if they are available.
		jtitle		Journal title. Use the most complete title available, e.g. "journal of the american medical association". Abbreviated titles, when known, are provided in the "stitle" element.
		stitle		Abbreviated or short journal title. This is used for journal title abbreviations, e.g. "J Am Med Assn
	date			The "date" element contains child elements that are used to express the fully qualified date, time, or season of when an individual article in a serial publication was published. The "date" element is not repeatable
		isoDate		Date of publication in ISO 8601 form YYYY, YYYY-MM or YYYY-MM-DD
		chron		Indications of chronology in a non ISO8601 form (like "Spring" or "1st quarter") should be carried in this element; the element content is not normalized. Where numeric ISO8601 dates are also available, they should be provided in the "date" element. As such, a recorded date of publication of "Spring, 1992" becomes "date=1992" and "chron=spring". Chronology information can also be provided in the "ssn" and "quarter" elements

Metasearch Search and Retrieval
Citation Level Data Elements

v 1.0

Element				Description
		season		Season (chronology). Legitimate values are "spring", "summer", "fall", "winter"
		quarter		Quarter (chronology). Legitimate values are "1", "2", "3", "4"
	context			The "context" element contains child elements that are used to express the fully qualified location of an individual article within a serial publication. This is the volume, issue, page number, etc. Since different publications may or may not have any one of these child elements, they are not specified here other than requiring a context element for citation level reference.
		volume		Volume designation. Volume is usually expressed as a number but could be roman numerals or non-numeric, e.g. "124", or "VI".4"
		part		Part can be a special subdivision of a volume or it can be the highest level division of the journal. Parts are often designated with letters or names, e.g. "B", "Supplement"
		issue		This is the designation of the published issue of a journal, corresponding to the actual physical piece in most cases. While usually numeric, it could be nonnumeric. Note that some publications use chronology in the place of enumeration, i.e. Spring, 1998.
		spage		Start, or first page number of a start/end (spage-epage) pair. Note that pages are not always numeric.
		epage		Second (ending) page number of a start/end (spage-epage) pair
		pages		Start and end pages in the form "startpage-endpage". This field can also be used for an unstructured pagination statement when data relating to pagination cannot be interpreted as a start-end pair, i.e. "A7, C4-9", "1-3, 6"
		artnum		Article number assigned by the publisher. Article numbers are often generated for publications that do not have usable pagination, in particular electronic journal articles, e.g. "unifi000000090". If article numbers are identifiers that follow a URI Scheme such as "info:doi/" the information should be provided in the Identifier Descriptor of the ContextObject, not in this "artnum" element. Likewise, if articles are identified by means of a registered URI Scheme such as the http scheme, the information should be provided in the Identifier Descriptor of the ContextObject
	id			The "id" element contains child elements that are used to describe the standard ID assigned to the journal, book, serial, etc. It may be the ISBN, ISSN, EISSN, CODEN, or SICI. Enumeration values: ISSN: International Standard Serial Number (ISSN). ISSN numbers may contain a hyphen, e.g. "1041-5653" EISSN: ISSN for electronic version of the journal. Although there is no distinction by format in the assignment of ISSNs, some bibliographic services now carry both the ISSN for the paper version and a separate ISSN for the electronic version. This data element is included here to allow expression of both types of ISSN numbers ISBN: International Standard Book Number (ISBN). The ISBN is usually presented as 9 digits plus a final check digit (which may be "X"), e.g. "057117678X". ISBN numbers may contain hyphens, e.g. "1-878067-73-7" SICI: Serial Item and Contribution Identifier (SICI)
	publisher			The name of the publisher is required, if available
	fultextURI			A URI link to the full-text article is required, if available.