

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 massaged 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