Z39.50 and Metadata

William E. Moen, University of North Texas

The Z39.50 protocol, specifically as defined in the Z39.50-1995 standard, addresses many of the problems of document discovery and retrieval and addresses several aspects of metadata, both implicitly and explicitly,

Of course, elements that are metadata from one point-of-view might not be metadata from another. For example, bibliographic elements (author, title, publisher, and hundreds more) are not metadata with respect to bibliographic records, though they may be metadata with respect to a document that a bibliographic record describes. The older (1992) version of Z39.50 had a bibliographic context, and so did not necessarily consider bibliographic elements to be metadata. (In other words, since metadata is, by some definition, data about object data, then if bibliographic elements are object data, by definition they are not metadata). Furthermore, there are elements which legitimately may be considered metadata, but which Z39.50 refers to differently, for example, as *variant information*. We mention this certainly not for the purpose of debating what is and is not metadata, but to emphasize that the Z39.50 community has been thinking about what has come to be know as metadata, perhaps without realizing it.

A fundamental premise of Z39.50 is the distinction between search elements and retrieval elements. These can coincide, but often they do not. Thus, a given element may be a search access point of a database record, but not a retrievable element, and vice versa. For example, a database record might be an image; the record may be searchable via a unique local identifier, but that identifier is not part of the image. Conversely a bibliographic record for a book might include a spine title, which is a retrievable element of the database record, but the record might not be searchable via the spine title.

Another Z39.50 premise recognizes that different metadata elements apply at different levels. Z39.50 defines the concept of a *document variant*. A given document may be available in various formats; the 'author' element, for example, would apply globally (independent of format) while the 'size' (and 'cost') would vary by format.

Various categories of data elements are supported by Z39.50, many of which can be considered metadata. Several are described below. Note that the first category pertains to searching and the remaining categories to retrieval, but elements in the latter categories may be included under the "search access points" category.

The following are among the categories of metadata directly supported by Z39.50.

- search access points — E.g. author, title, publisher. In Z39.50 terminology, these are *search attributes*.
- format — Compression; body part type, e.g. ascii, postScript, SGML.
- formatting — E.g. lines per page, characters per line, columns, dots per inch, paper size, font.
- document descriptor — E.g. A DTD for an SGML document.
- language/character set — The language and/or character set of a document variant or a retrieved document.
- cost/size — The cost to retrieve a document (or a document variant), and its estimated size.
- surrogate relationships — A client may have retrieved a thumbnail of, say, an image, and the server might provide a pointer to an image for which the thumbnail is a surrogate.
- satisfying portions — Included with a retrieved document (or in lieu of the document) may be ranked *hit vectors*, each pointing to a section bearing some relationship to the search.
- Series order — For hierarchical data, this indicates how child nodes are ordered. E.g. 'sequential' (by pages, frame, screen), 'chronological', 'semantic size' (e.g. increasingly comprehensive abstracts), 'generality' (e.g., thesaurus words, increasing generality), 'concentricity'.
- integrity — This provides a measure of the server's estimation of authenticity of a derived object, relative to the original.
- processing instructions — The server may include, along with a document, machine-processible instructions on how to display the document to the user.
- restrictions — Pertaining to copyright and redistribution.
- user message — A message, included along with a document, that the server request the client to display to the user.
- score/rank — A rank, or normalized score, of a document, relative to other documents in the result set.