September 1999

We Used to Call it Publishing-Issues in Crosswalking Content Metadata Standards

Margaret St. Pierre

Blue Angel Technologies, Inc.

Sandra K. Paul

SKP Associates, SKPAssoc@cwixmail.com

Albert Simmonds

SKP Associates, awsimmo@ibm.net

Bill LaPlante

Bureau of the Census

Follow this and additional works at: <https://docs.lib.purdue.edu/atg>

Part of the [Library and Information Science Commons](#)

Recommended Citation

St. Pierre, Margaret; Paul, Sandra K.; Simmonds, Albert; and LaPlante, Bill (1999) "We Used to Call it Publishing-Issues in Crosswalking Content Metadata Standards," *Against the Grain*: Vol. 11: Iss. 4, Article 31.

DOI: <https://doi.org/10.7771/2380-176X.3551>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

Biz of Acq
from page 72

public services staff was that we should be able to search for "Forms on disk," and retrieve all books with forms on disk. This could be achieved by using the "Added title" field. Of course, this suggestion gave our cataloging librarian fits. "Forms on disk," she told me, "must appear somewhere for it to be entered into the added title field. Our rationale for creating this added title cannot be because it would be convenient. It runs entirely against AACR2 rules ... we can't do it." The cataloging librarian's solution was to put "forms on disk" into a note field which could be indexed and searched using the keyword index. It would involve creating an indexed note field for each book with forms on disk. We would be able to do a keyword search for "forms on disk," and limit the results by subject, though we would not be able to go the other way, i.e., we would not be able to do a subject search and limit it to forms on disk. The solution was inelegant, but workable.

The final answer came at a small conference for local acquisitions librarians who use the III system. One of the librarians present said that her library was using the "alternative title field" to index familiar names for books. Innovative will index this field—793 in the MARC record—as an added title. On the staff display it shows up as an added title; it does not show up at all on the public display. Significant for our purposes, entries in this field would not be governed by stringent AACR2 convention. "Alternative titles" do not have to appear anywhere on the item in question; they can be made up "out of whole cloth." I realized that this was the answer to our problems. We could use this field to enter the title "forms on disk."

We have gone on to use the 793 field to index all sorts of non-AACR2 language. Most recently, one of our staff has begun indexing books by color. This will eventually become a clickable page on our WebPAC—handy for such questions as: "I'm looking for the contracts casebook—you know, the green one." Subject: contracts, Limit by: color. The point of the study, though, is not the success, but the process. Having someone familiar with the workings of both departments to serve as an intermediary was essential to solving the problem. The reference department says: "Here is the problem (cannot search for forms on disk) and here is what we want (added title access)." The cataloging department says: "Here is why we cannot do that (AACR2)." In many institutions, the discussion might end there; the adversarial relationship between public and technical services would be preserved. A person who works for both departments, however, can better negotiate between the two. In our library that meant that I could go back to our cataloging librarian and say, "I understand that AACR2 doesn't allow us to do this. Is there another indexed or indexable field we could put this in?" And when we came up

with a workable but inelegant solution, I was in the best position to keep looking for a better one.

Conclusion

Using the case study above, I have tried to show one advantage of having technical services staff work in reference and/or public services. There are many others. Working on the front lines of public service, I hear the praise of our library, of our collection—praise which in large part belongs to the technical services staff. From the distant land of the back office, it's easy to forget: patrons do very much appreciate what we do. Also, for those of us who spend the days acquiring and organizing information, it is very helpful to know how people ask for information. Remember, the entire case study above was borne of a reference question: "Do you have any bankruptcy books that are accompanied by forms on disk?"

Automated library systems have started us down the path of integrating public and technical services. As libraries grow more electronic, that integration will grow even more essential. What good to a library is a technically dazzling

BOOKS & PERIODICALS

ONLINE

a directory of online publications

JUST A
CLICK AWAY!

editor • Nuchine Nobari

HOW LONG DOES IT TAKE YOU TO FIND
WHAT YOU'RE LOOKING FOR ONLINE?

- 97,000+ Sources Online
- 14,000+ Fulltext
- Links to 3,000+ URLs, all verified

AVAILABLE AT
WWW.PERIODICALS.NET

FULLTEXT • ABSTRACTS • CITATIONS

LIBRARY TECHNOLOGY ALLIANCE, LTD.
POST OFFICE BOX 77232 (202) 789-2099 PHONE
WASHINGTON, DC 20013 (202) 789-2474 FAX

system if it isn't designed with an eye toward public use? Of what use is an ingeniously scripted Web page if it isn't designed with an eye toward the ways in which people use information? I suspect that the division in libraries between technical and public services will endure for some time to come. For us to grow with the world around us, the stark line between the two must begin to blur.

Used to Call it Publishing Endnotes from pg 76

¹ LaPlant, W. and St. Pierre, M. Taxonomy of Metadata Standards. Copies are available from the U.S. Bureau of the Census, Statistical Research Division, Attn: William LaPlant, Room 3000 FB 4, Washington, DC 20233-9100.

² Dublin Core: <<http://purl.oclc.org/dc/>>.

³ USMARC: <<http://www.loc.gov/marc/>>.

⁴ Federal Geographic Data Committee (FGDC): <http://www.census.gov/geo/www/standards/scdd/CDsupplement.html>.

⁵ Government Information Locator Service (GILS): <http://www.usgs.gov/gils/prof_v2.html>.

⁶ Directory Interchange Format (DIF). <<http://gcmd.gsfc.nasa.gov/difguide/difman.html>>.

⁷ Inter-University Consortium for Political and Social Research (ICPSR) Data Documentation Initiative (DDI): <<http://www.icpsr.umich.edu/DDI>>.

⁸ Survey Design and Statistical Methodology (SDSM). Bureau of the Census Draft Standard. Copies available from the U.S. Bureau of the Census, Statistical Research Division, Attn: SDSM, Room 3000 FB 4, Washington, DC 20233-9100.

⁹ Gillman, D.; M. Appel; and S. Highsmith. Building A Statistical Metadata Repository, Second IEEE Metadata Conference 1997.

¹⁰ LaPlant, W. P. Jr., Lestina, G. J. Jr., Gillman, D. W., and Appel, M. V. (1996), *Proposal for a Statistical Metadata Standard*, Census Annual Research Conference, Arlington, VA, March 18-21, 1996.

¹¹ Consortium for the Computer Interchange of Museum Information (CIMI): <http://www.cimi.org/downloads/CIMI_profile/profile.htm>.

¹² Information Resource Dictionary System (IRDS) Content Model Standard: <<http://www.irds.org/gradwellworkshoppaper.html>>.

¹³ Murata, M. DTD Transformation by Patterns and Contextual Conditions, GCA SGML/XML 1997, pp. 325-332.

Element 1 In GILS, Contact Name is subordinate to Point of Contact.

Element 2 In DIF, the First Name and Last Name elements are subordinate to the Technical Contact element.

We Used to Call it Publishing — Issues in Crosswalking Content Metadata Standards

by **Margaret St. Pierre** (Blue Angel Technologies, Inc.) and **Bill LaPlante** (Bureau of the Census)

Column Editors: **Sandra K. Paul** (President, SKP Associates) <skpassoc@CWIXmail.com> <sandy@bookinfo.org> and **Albert W. Simmonds** (SKP Associates) <awsimmo@ibm.net>

Metadata is one of the key links required for search and discovery. What is required to harness the variety of metadata sets that are beginning to proliferate? The authors explore some of the issues that surface in considering the harmonization of metadata standards. This paper originally appeared on the NISO Website as the **First NISO White Paper**. The authors extend their thanks to **Jim Restivo** of Blue Angel Technologies for his help in illuminating key ideas and reviewing early drafts.

Introduction

This paper describes the general issues involved in the harmonization of metadata standards and in the development of crosswalks between related metadata standards. We begin by enumerating a set of simple procedures for harmonizing metadata standards. Next, we describe the set of criteria needed to develop a fully specified crosswalk. Finally, we propose future steps for simplifying crosswalk implementation through the use of formal specifications and automation.

A content metadata standard is an open specification that itemizes a set of elements and their meanings.¹ Each element is tagged with an identifier (e.g., Title, Author) that distinguishes the element from other elements within the standard. Each element has a set of constraints or rules specifying the allowable content of the element and its relationship to other elements within the standard. For example, a Country element may be restricted to a standard vocabulary of country codes, and may have a subordinate relationship to an Address element. A content metadata standard, hereafter referred to as a metadata stan-

dard, is typically developed to support a specific community of interest. A large number of metadata standards have been developed, and many more are underway. Examples of popular standards include Dublin Core², USMARC³, Federal Geographic Data Committee (FGDC)⁴, Global/Government Information Locator Service (GILS)⁵, Directory Interchange Format (DIF)⁶, Inter-University Consortium for Political and Social Research (ICPSR)⁷, Survey Design and Statistical Methodology (SDSM)^{8,9,10}, Consortium for the Computer Interchange of Museum Information (CIMI)¹¹, and the Information Resource Dictionary System (IRDS) Content Model Standard¹².

To reach the broadest community of users, information must be made available in accordance with a number of related metadata standards. As the number, size, and complexity of metadata standards continues to grow, supplying the metadata for each standard becomes more and more time-consuming and tedious. In order to minimize the amount of time needed to create and maintain the metadata, and to maximize its usefulness to the widest community of users, there is a mounting need for the metadata maintained in one standard to be accessible via alternate standards. A crosswalk is a specification for mapping one metadata standard to another. Crosswalks provide the ability to make the contents of elements defined in one metadata standard available to communities using related metadata standards. The specification of a crosswalk is a difficult and error-prone task requiring in-depth knowledge and specialized expertise in the associated metadata standards. Obtaining the expertise to develop a crosswalk is particularly problematic because the metadata standards themselves are often developed independently, and specified differently using specialized terminology, methods and processes. Furthermore, maintaining the crosswalk as the metadata standards change becomes even more problematic due to the need to sustain a historical perspective and ongoing expertise in the associated standards.

Harmonization is the process of enabling consistency across metadata standards. Harmonization of metadata standards is essential to the successful development of crosswalks between metadata standards. Harmonization results in the ability to create and maintain only one set of metadata, and to map the metadata to any number of related

metadata standards. The use of harmonization vastly simplifies the development, implementation and deployment of related metadata standards through the use of common terminology, methods and processes.

Harmonization

The first step toward harmonization is to extract the common terminology, properties, organization, and processes used by many of the metadata standards, and create a generic framework in which to develop new or revise existing metadata standards. Because similar procedures can be applied to related metadata standards, the implementation of the standards and the development of new crosswalks are simplified. This section highlights several key steps for obtaining harmonization among metadata standards.

Terminology

A lack of common terminology currently exists among the different metadata standards. For example, USMARC metadata is identified using the terms: "tag," "indicator" and "subfield code," whereas Dublin Core identifies its metadata using the term "label." As a starting point for harmonization, it is essential to agree on a common set of terminology to be used in the specification of metadata standards. As part of the task of identifying common terminology, it is crucial to also establish a formal definition for each term. A shared vocabulary prevents misinterpretation of the standards, and lays the foundation for subsequent harmonization efforts.

Properties

Many of the metadata standards use similar properties in the definition of their metadata. The similarities need to be extracted and the concepts generalized and used in a common way across all metadata standards. Examples of common metadata properties include: a unique identifier for each metadata element, e.g., tag, label, identifier, field name; a semantic definition of each metadata element; whether or not a metadata element is mandatory, optional, or mandatory based on certain conditions; whether or not a metadata element may occur multiple times; the organization of metadata elements relative to each other, e.g., hierarchical, parent-child relationships; constraints imposed on the value of the element (e.g., free text, numeric range, date, or a controlled vocabulary); optional support for locally defined metadata elements. The

continued on page 75

We Used to Call it Publishing
from page 74

shared properties can then be expressed and used in a similar fashion within each metadata standard. This step is important for simplifying crosswalk development.

Organization

The development of a crosswalk is often complicated by the fact that each metadata standard is organized differently. To ease the ability to find information within any given metadata standard, each document should be organized in a similar manner, so that a given section of one standard can be found in an analogous section of another standard. The generic metadata document organization should also accommodate the content-specific requirements of different metadata standards.

Process

During the development of a metadata standard, there may be occasions where the choice of mechanism or process selected for use in the metadata standard is arbitrary, and there is an analogous process in another related standard. Harmonization is achieved when the analogous processes are chosen to be the same. The primary benefit of unifying the selection process is in simplifying crosswalk development. For example, because CIMI metadata incorporates the elements defined by Dublin Core metadata, in addition to a large number of content-specific museum metadata, developing a crosswalk between the two metadata standards is simplified.

Semantic Mappings

Foremost in the harmonization effort and eventual crosswalk development is the intellectual task of determining the semantic mapping of elements between the source and target metadata standards. The task involves specifying a mapping of each element in the source metadata standard with a semantically equivalent element in the target metadata standard. The prerequisite to a meaningful mapping requires a clear and precise definition of the elements in each standard. Many metadata standards already provide a semantic mapping to related metadata standards. These mappings are often tabulated in informative appendices to the metadata standard. Although a semantic mapping is an essential piece of the crosswalk, a number of additional issues must also be resolved to obtain a complete crosswalk. These additional components are discussed in the following section.

Crosswalks

A crosswalk is a set of transformations applied to the content of elements in a source metadata standard that result in the storage of appropriately modified content in the analogous elements of a target metadata standard. A complete or fully specified crosswalk consists of both a semantic mapping and a metadata conversion specification. The metadata conversion specification contains

the transformations required to convert the metadata record content compliant to a source metadata standard into a record whose contents are compliant with a target metadata standard. A fully specified crosswalk requires that all implementations of the crosswalk on a specific source content result in the same target content. If two different implementations of a crosswalk operating on the same source content result in different target content, the crosswalk is not fully specified. This section describes the metadata conversions that must be addressed in a fully specified crosswalk.

Element to Element Mapping

All metadata standards specify a number of properties associated with the specification of the various metadata elements. Some standards, such as USMARC, qualify each element as repeatable or non-repeatable. Some metadata standards indicate whether or not an element is mandatory or optional. Others, such as FGDC, incorporate both these attributes into a single property by indicating a lower and upper bound on the number of times an element may occur. An inclusive lower bound of zero indicates an optional element, whereas an inclusive lower bound of one indicates that the element must occur at least once and thus is mandatory. For crosswalk development, these properties must be taken into consideration for the mapping of each element. The trivial case is mapping elements that have identical properties, e.g., mapping mandatory, non-repeatable elements to mandatory, non-repeatable elements of identical data content types. There are a number of interesting cases that require more complex resolution.

"One to Many"

In most cases, a one-to-many map is trivial; an occurrence of the source element maps to a single occurrence in the target element. There are cases where the mapping requires more explicit resolution. For example, the source standard may contain a non-repeatable keyword element. The element definition specifies that its element value is made up of one or more keywords separated by a semicolon character. This element may map to a repeatable element in the target standard where each keyword must occur as a repeated element. In this case, the mapping requires specialized knowledge of the composition of the source element, and how it expands into multiple target elements. Another interesting case is that of mapping one source element to two unique target elements. For example, a crosswalk for GILS to DIF would need to map the GILS Contact Name element1 to the First Name and Last Name elements2 in DIF. In this case, general rules must be specified to correctly extract the first and last name from the GILS element and map them to the corresponding DIF elements.

"Many to One"

The many-to-one map must specify
continued on page 76

Commonsense Copyright

A Guide for Educators and Librarians, 2d ed.

by R.S. Talab

\$39.95 304 pages, sewn soft-cover, notes, appendices, charts, resource guides, checklists, index
ISBN 0-7864-0675-5 1999

As it applies to electronic publishing, permissions, library materials, schools, the Internet, and more.

Internet Research

Theory and Practice

by Ned L. Fielden
and Maria Garrido

\$25 181 pages, sewn softcover, tables, illustrations, appendix, bibliography, index
ISBN 0-7864-0538-4 1998

"Fills an important niche"
— *Library Journal*

"Just about every subject a would-be researcher could need"
— *Public Libraries*

Angels

An Indexed and Partially Annotated Bibliography of Over 4300 Scholarly Books and Articles Since the 7th Century B.C.

by George J. Marshall

\$95 library binding (7 × 10)
485 pages, indexes
ISBN 0-7864-0555-4 1999

"A serious study ... useful"
— *Against the Grain*

McFARLAND

Box 611, Jefferson NC 28640
Orders 800-253-2187
FAX 336-246-5018
www.mcfarlandpub.com

what to do with the extra elements. If the resolution is to map all values of the source element to a single value in the target element, explicit rules are required to specify how the values will be appended together. Alternatively, if the resolution is to only map one source element value to the target, with the possible consequence of information loss, the resolution must indicate the criteria for element selection, e.g., the first element, or the most recently added element. Another important case that requires resolution is the handling of a source element that does not map to any appropriate element in the target standard. Since many metadata standards provide the ability to capture additional information, the resolution must specify precisely how the element value is to be added. In some cases, there may be mandatory elements in the target that have no corresponding mapping in the source metadata standard. Because the target requires a value for the mandatory elements, the crosswalk must provide a resolution for their values.

Hierarchy, Object, and Logical View Resolution

The crosswalk must address how to resolve differences in the hierarchy, object, and logical view orientation of the different metadata standards. Most metadata standards organize their metadata hierarchically. In some cases, the depth of the hierarchy may be fixed, as in the USMARC and GILS standards. In other cases, standards such as SDSM and FGDC are recursively defined where the depth of the hierarchy is unlimited. A few standards, such as SDSM and IRDS, are multiple object metadata standards. Multiple object metadata means that the metadata concerns more than one item. For example, USMARC, despite its size and complexity, is a single object metadata standard, in that the metadata is always associated with only one item per use. FGDC, CIMI, and ICPSR are all single object metadata standards, specifying the metadata associated with a geospatially referenced dataset, physical artifacts or their electronic derivatives, and the data resulting from political or social research, respectively. On the other hand, in both IRDS and SDSM, the metadata for many objects may be associated together in various ways depending on the perspective with which the metadata is being retrieved. IRDS objects include system, program, and data, for example, while in the SDSM some of the objects include survey, documentation, dataset, and frame. In a multiple object metadata standard, the hierarchy for retrieving or manipulating the metadata is dependent on the object of interest. Standards

such as SDSM and IRDS also specify multiple logical views of their metadata elements. A logical view enables users to see a specific set of metadata elements of the metadata standard organized in a specific way. The standards that provide multiple logical views enable different user communities access to the same metadata elements using different organizations, hierarchy, or representations of those metadata elements.

Content Conversion

Metadata standards typically restrict the contents of each metadata element to a particular data type, range of values, or controlled vocabulary. Conversions are required between text and numeric values or text and date values. Often the needed conversions are based not only on the defining properties of the source and target metadata elements, but also on the contents of the source metadata elements. Resolution rules are required between a source element whose value is specified as free text and a target element whose value is restricted to a controlled vocabulary. Resolution rules are also necessary, for example, between a source and target element using different controlled vocabularies.

Conversion Combinations

When conversion properties are considered independently, the metadata conversions may appear to be straightforward to specify and process. In practice, however, several conversion issues often surface in combination, which significantly complicate the conversion specification and process. For example, converting a source metadata element that is both hierarchical and repeatable to a target metadata element that is not repeatable and does not share the same hierarchy is not a straightforward process. Consideration must be given to the transformations needed for converting to a target metadata element where multiple properties are substantially different from the source.

Future Steps

If several metadata standards were harmonized and a fully specified crosswalk between related standards was developed, the next step would be to work toward the goal of automating the crosswalk process. Toward this goal, if the metadata and crosswalk transformations could be captured in a formal way that is consistent throughout the many metadata standards, the implementation of the standards and their crosswalks would be vastly simplified. Fully automated crosswalks would also enable search engines to function with any given family of metadata standards. This section proposes the idea of generalizing and formalizing metadata and their crosswalks.

Formal Content Metadata Specification

There is currently no common established means for specifying metadata and

its associated properties. Most metadata standards use an arbitrary combination of free text descriptions, lists and tables. More complex metadata standards, such as FGDC, use a set of production rules to specify their metadata properties. It may be possible to generalize and formalize the specification of metadata using a canonical representation or metadata specification language (MSL). This procedure is analogous to specifying the syntax of a programming language using the popular Backus-Naur Form (BNF). The purpose of an MSL would be to establish a consistent means for specifying metadata and its many properties. The result of using a generic MSL for all metadata standards is simplification of the metadata specification process and attainment of a concise and more precise representation of each metadata standard. Like BNF, the use of an MSL would not capture the semantics of the metadata elements. A straight syntactic description of a given metadata standard is inadequate for capturing all the information needed to automate a crosswalk. A minimum set of data types would also need to be defined. This minimum set would be used to derive all other data types needed to represent metadata elements in a target metadata standard.

Formal Crosswalk Specification

Like content metadata specification, there is no standard method of specifying all the transformations required for crosswalk development. Most of the crosswalks that are provided with the metadata standards are no more than a semantic mapping of elements. A fully specified crosswalk must also provide a metadata conversion specification, which includes rules for element to element mappings, hierarchy and object resolution, and metadata content conversions. To automate the implementation of a crosswalk, it may be possible to formalize the specification of a crosswalk. Research has shown that it is possible to formally specify a set of transformations to convert a source metadata standard to a target standard using the theory of tree automata. For example, if the source and target metadata standards are specified using a Standard Generalized Markup Language (SGML) Document Type Definition (DTD), the crosswalk could specify a set of transformations for converting the source DTD to the target DTD¹³.

*NB: This article first appeared as a NISO White Paper; it is posted on the NISO Website (www.niso.org). It also was published in NISO's newsletter *Information Standards Quarterly*, January 1999 (v.11, #1) and is reprinted here with the permission of NISO Press.—KS*

See Endnotes on page 73