

CASE 8

Sport Schema Crosswalks: Appraising, Converting, and Preserving Proprietary Schema to Open Source Standards

AUTHOR: **JAMES W. ROGERS**
Digital Initiative Library Team Intern
Penrose Library/Digital Initiative/Special Collections and Archives
University of Denver (wolfwerks@mac.com)

PAPER DATE: July 2008

CASE STUDY DATE: Ongoing

ISSUE: Building a model for crosswalking or mapping of a proprietary XML schema to an open source XML schema for long-term archival preservation. The original data was not in an archival standard format and the goal of the department is to preserve sports data and provide access; therefore we proposed using the open source SportsML schema to preserve the data.

KEYWORDS: Appraisal issues, Data format issues, Data integrity issues, Data longevity issues, File format issues, Metadata Standards

Copyright by James W. Rogers.

Background

This case study discusses and proposes a model for crosswalking, or mapping of a proprietary Extensible Markup Language (XML) schema to an open source XML schema for long-term archival preservation. The proprietary schema is encoded in MS-DOS with a generated output in XML and is part of a game-scoring and statistics software package used by the University of Denver's athletic department. This data is not in an archival standard format and the goal of the department is to preserve sports data and provide access; therefore we propose using the SportsML schema to preserve the data. SportsML is an open source extensible language that will provide a stable archival standard. It is envisioned that a crosswalk will be built that maps the different elements and attributes of the two schemas. An Extensible Stylesheet Language (XSL) will be incorporated to display output. The best way to proceed with this complex undertaking, as well as the potential problems and pitfalls, are issues that must be addressed by the university digitization team in order to prove that this potential concept will be viable.

Institutional Context

The University of Denver (DU) was founded in 1864 as Colorado Seminary by John Evans, who was appointed by Abraham Lincoln as the second governor of the Colorado Territory. DU is the oldest private university in the Rocky Mountain Region. It is a coeducational, four-year university located in Denver, Colorado, with an approximate student body of 11,000 enrolled in various graduate and undergraduate programs. Students at DU hail from 78 countries and are enrolled in 16 schools and colleges within the university. DU employs 484 full-time faculty, 93 percent of whom hold either doctoral or the highest degree appropriate to their discipline; DU also employs 41 part-time faculty and 473 adjunct faculty. Notable alumni include former Interior Secretary Gale Norton, Secretary of State Condoleezza Rice, Secretary of Veterans Affairs James Nicholson, U.S. Army Gen. George Casey, Jr., and Andrew Rosenthal, the assistant managing editor of *The New York Times*. In the field of athletics, the University of Denver's Division of Athletics and Recreation sponsors 17 intercollegiate varsity sports competing as members of National Collegiate Athletic Association Division I.

The University of Denver's Penrose Library has more than 3.1 million volumes, including 6,607 periodicals. The Department of Special Collections and Archives at Penrose contains the university's Archives, Rare Books, and Manuscripts; the Carson-Brierly Dance Library; and the Ira M. Beck Memorial Archives. In addition, the digitization projects of the Special Collections and Archives Department are multifaceted endeavors to digitize and catalog materials from collections that reside in the department. Currently the Division of Athletics and Recreation has partnered with the library to bring forward the rich documentary history of athletics and recreation at DU.

There is a growing online digital selection of historical photographs, team rosters, and the University of Denver yearbooks, the *Kynewisbok*. As the collaborative digitization project continues, more photographs, rosters, and yearbooks will become available

online, as well as press releases, team programs, audio, video, and statistical sports information.

Introduction

The University of Denver's (DU) athletic department currently uses a proprietary software package called Stat Crew that allows the department to capture game, player scores, and statistics for the various sports that the school participates in. The software captures and displays various data including, but not limited to, box scores, play-by-play reports, and press releases. It also allows the department to capture single-game statistical information and accumulate season-to-date statistics for most major sports. This software is used by many college and university teams across the country and is one of the most popular sports scoring and statistical software tools available. The software engine behind Stat Crew is based on MS-DOS and game files are generated with proprietary VPK extensions. The software also allows users to generate XML data that may be used to report game statistics electronically to such organizations as the National Collegiate Athletic Association (NCAA). Stat Crew also generates HyperText Markup Language (HTML) output for web display.

Since Stat Crew is proprietary software, the strong potential exists that data, information, and records captured by DU will not be preserved for the long term. The company itself may no longer be a viable entity, or the information that is captured by the software may no longer be readable or accessible. In other words, the Stat Crew data is not in an archival preservation format and the athletic department strongly desires to have its athletic records permanently preserved. While the XML output generated by Stat Crew in itself is a viable language and can be preserved, the information and the format that it presents is proprietary and cannot easily be used in other programs or applications, much less displayed simply and openly in public repositories.

The Digital Library Team (DLT) at DU's Penrose Library, in cooperation with DU's athletic department, was charged with providing a solution to this problem. The skill sets of the DLT for this specific project consisted of some minor familiarity with XML parsing, and Oxygen XML editor software used as a primary tool, as well as minimal knowledge of XSL transformation. The team did not have a firm grasp of crosswalk and mapping concepts and began by deliberating the best way to initiate the project. We established the initial goals of the of the Stat Crew crosswalk project as follows:

1. Appraise the Stat Crew data and records.
2. Normalize the XML data.
3. Convert or translate Stat Crew XML records and data to Text Encoding Initiative (TEI) XML records and data.
4. Catalog the output to our Rediscovery catalog system.
5. Export the translated archival sport data and records to the Colorado Alliance of Research Libraries (ADR) repository in Denver, Colorado.

The ADR is a consortium of nine institutions representing eleven major libraries located in Colorado and Wyoming.

Discussion of our goals, problems, potential roadblocks, and ongoing changes in our processes follow below.

Appraisal

Several meetings were held between the digital library staff and the athletic department in order to determine the types of records and statistics that should be transformed. Since DU has a very strong history and heritage involving ice hockey, we decided to experiment with just one type of record in attempting to translate the schema—a hockey game played between North Dakota and the DU Pioneers in November 2007. It should be noted that, in using an appraisal approach, communication between the athletic department and the DLT proved paramount. We realized that, as work continued on the project, the cost and time spent would increase in proportion to the amount of data and records that would be transformed into SportML. Using every statistic and record within a Stat Crew game and converting them into the SportML schema would involve complex XSL transformations and would also involve substantial programming time. Therefore, it became necessary to determine the minimum amount of information that the athletic department required. If too much information was transformed, the project might become bogged down in wasted time and funding. If too little information was transformed, however, the records might prove useless by not providing the user with the desired or specific amount of information. This point must be emphasized since the DLT has a limited amount of staff, funds, and priorities. Funding for the project is provided by the DU athletic department, and is just one component of a yearly grant to the Penrose DLT. Funding for the current year (2008) is approximately \$50,000. Currently the DLT is negotiating with the athletic department to determine its needs and the exact amount of information and records that will be transformed.

Normalization

In attempting to normalize our XML data, we came upon various definitions of normalization. The *IBM Glossary of Unicode Terms* (2008) states that “normalization is the process of converting Unicode text into one of several standardized forms in which precomposed and combining characters are used consistently.” The *DMReview* (2008) glossary states “normalization is the process of reducing a complex data structure into its simplest, most stable structure. In general, the process entails the removal of redundant attributes, keys and relationships from a conceptual data model.” We believe that this definition offers a logical concept towards achieving our goal and will discuss this further below. In addition, normalization also should include the adoption of basic guidelines that realize our goals of normalizing XML, such as:

- Eliminating ambiguity in data expression.
- Minimizing redundancy.

- Facilitating the preservation of data consistency.
- Allowing for rational maintenance of data.

At the very minimum, our normalization should incorporate the process of converting complex data structures into simple and stable data structures. Normalization of the Stat Crew data will mean that the information that Stat Crew contains is readable and convertible into a standard that is consistent and stable, thus allowing the user to readily access records and other data.

Conversion or Translation

One of the largest obstacles to successful completion of our project was the conversion of Stat Crew XML to a TEI XML format. Since TEI was developed to provide guidelines for the preparation and interchange of electronic texts for scholarly research, we realized quickly that TEI would not be appropriate for use with the Stat Crew XML schema. TEI would not suitably translate sports statistics and records into a format that would easily translate to a new XML archival format. After much research, we decided that SportsML would be a superior fit for our needs and give us additional options that TEI could not. It also became obvious that conversion of the schemas and the crosswalk itself is a custom design and not built to any standards. In addition, the mappings initially may be imperfect and the crosswalk may evolve and change because of this lack of standardization. Owing to this lack of a specific standard, interoperability with other repositories may eventually present some additional problems that will require resolution.

SportsML

SportsML is an open source global XML standard for the interchange of sports data. SportsML uses XML to describe the content and structure of sports data and statistics. SportsML has the potential to make it easier to integrate sports feeds and data that adhere to SportsML than to rely on other proprietary formats. SportsML was launched in March 2001 by the International Press Telecommunications Council (IPTC) as part of an endeavor to create, track and follow specialized vocabularies for data that interest the news industry. The current release of SportsML is version 1.8, with a newer version, 2.0, scheduled for tentative release in the latter part of 2008 when the IPTC will vote on the 2.0 package. SportsML now supports the description and identification of many sports characteristics including scores, schedules, standings, statistics and news. SportsML contains a core Document Type Definition (DTD) that uses properties to describe a wide range of sports and sport coverage. SportsML also has several "plug-in" specific-sport DTDs, which can be used to provide in-depth data for a specific sport such as ice hockey. Although SportsML currently covers a limited amount of sports, this does not imply that it is limited in its future coverage and growth. The core DTD allows the development of additional plug-ins for additional sports. In fact, SportsML encourages users, developers and other interested parties to contribute to its development (SportsML.org, 2008).

Crosswalk/Mapping

Once we decided that SportsML might meet our needs, we had to decide how a crosswalk or mapping of the two schemas might be accomplished and whether or not it was possible. There were three components that we had to consider and develop: the source XML metadata, the crosswalk, and the target metadata. We realized that, for the crosswalk or mapping of the metadata schema to work, we had to thoroughly map all of the elements and attributes of both of the schemas. Stat Crew elements would have to be matched with SportsML elements and Stat Crew attributes would have to be matched to SportsML attributes. The easiest way to do this was to manually input all of the different elements and attributes into a Microsoft Excel spreadsheet. While this proved to be a very time consuming procedure, the more difficult aspect of this work involved the fact that the elements and attributes for the schemas did not easily or accurately match. In other words, the concepts or equivalences of each of the schemas are slightly different and some ambiguity exists between them. As Example 1 demonstrates, there is not a one-on-one element or attribute match for each schema.

Example 1

Stat Crew:

SportsML:

<u>Element</u>	<u>Attribute</u>	<u>Element</u>	<u>Attribute</u>	
venue	gamid	event-metadata	team=, id= full= team=, id= full= date-coverage-type= site-city=, site-state=	
	visid			
	visname			
	homeid			
	homename			
	date			
stars	location	standing-metadata	content-label=	
				star
				star
status	period	name	uniform-number full= period-value=	
				vh
				uni

By contacting Stat Crew's vendor support we were able to obtain an ice hockey game XML description that gave us tag definitions for the various attributes and elements, which helped in our mapping of the Stat Crew schema. Other descriptions of various sports were available, but since we chose ice hockey as our first experimental venture we did not need those for this pilot project. SportsML has an online version, 1.8, of its element and attribute concepts in addition to a printable version, which we found readily accessible.

As our mapping of the elements and attributes continued, we consulted with the Penrose Library IT team and found that an XSLT could be written for the mapped data. A precise appraisal that delineated the specific records and essential information that the department wanted to capture, however, needed to be conducted before any XSLT should be attempted. Since both the DLT and the IT team had limited resources and time, and would have to spend considerable effort on a custom application that would perform the transformation, it was essential to find out exactly what would be the minimum amount of information and data to be transformed. Examples of minimum essential data for an ice hockey game would be game date, city location, game venue, team names, player names, periods, scores during periods, and final scores.

Cataloging and Exporting

Our ultimate goal involved converting the Stat Crew schema to SportsML and exporting the data to the ADR, where it would be converted to HTML for user access. Once the XSLT phase of our project is complete, the output will be cataloged into the library's Rediscovery catalog system and imported into the ADR. ADR uses Fedora, an open source, general-purpose digital object repository system. Fedora uses a simple XML format called FOXML, which expresses the Fedora digital object model. With the current 2.0 version of Fedora, digital objects such as the proposed Stat Crew/SportsML records and documents will be stored internally at the ADR in the FOXML format. Fedora also supports the continuation of Metadata Encoding and Transmission Standard (METS) in importing and exporting objects to and from the ADR consortium (Fedora, 2008). The DPT may find it necessary to use METS or Metadata Object Description Schema (MODS) to import the transformed schemas into the ADR.

Mapping Software

Several "off-the-shelf" software programs presented themselves as a possible viable solution to the mapping problem. One, by Stylus Studio, was an XML-to-XML mapper that might work once all of the elements and attributes had been fully mapped out. Stylus Studio is relatively inexpensive (\$350–\$600, depending on the version) and offers a tool that may work for us in the future. Another inexpensive software package (\$140) by Rustemsoft, the "XML Converter Professional Edition" also provides a means of data transformation and conversion. Both of these software tools used visual mapping and had a user-friendly graphical mode. Finally, we downloaded Allora Software's XML mapping and database program. This tool provides developers with bi-directional access to relational databases without the need for XSLT programming. It transforms data structures between XML elements and attributes, and between database structures. Since our data was not maintained in a database, we decided that this package would not meet our needs. In addition, Allora software's license price exceeded our budget. We finally decided to proceed without purchasing any additional software, using our own resources for the moment and possibly revisiting the purchase of a mapping software tool at a later date.

Recommendations

The DLT recommended the following steps to complete the project.

1. Transfer all of the Stat Crew XML records, scores and statistics from all sports that DU participates in onto a secure server, preferably in the library itself. The athletic department will have access to the records, but the DLT and the library will have ultimate responsibility to maintain the Stat Crew data. Full archival backup standards and criteria will be met with either off-site back up of the Stat Crew data or by backing up the data to another format such as DVDs.
2. Discuss and decide in collaboration with the athletic department the exact records, scores and statistics that will ultimately need to be preserved and archived for future cataloging in the Rediscovery system and for access by ADR participants. This will possibly take some time and effort and will be determined by the amount of funds and resources necessary in order to translate the schemas. It must again be emphasized that excellent archival appraisal decisions at this point will help to determine the overall success or possible failure of the project.
3. Once the appraisal phase is completed, full mapping of the schemas can continue or begin for new sports. Since this is a time consuming effort, the possibility exists that work-study undergraduate or graduate students may be trained to convert the element and attribute sets from Stat Crew to SportsML. A short instructional course or workshop on XML, schemas and metadata will be necessary for the students in order to minimize confusion and so that they understand the concepts of mapping or crosswalks between schemas. It might be desirable to only map one sport, such as ice hockey, in the initial phase.
4. After a map or crosswalk has been completed, the IT team can develop an XSL to transform the schemas. The IT team will already have participated in the appraisal phase and will provide input into the amount of time and resources that would be devoted to developing or writing the XSLT. It is imperative that the IT team be involved closely with the ongoing developments and decisions that the DLT undertakes. Also at this time, consideration may be given to purchase or license one of the previously discussed mapping software programs, such as the packages available from Stylus Studio and Rustemsoft.
5. Once a successful XSLT is completed, the resultant records can be cataloged in the Penrose Library Rediscovery database and become accessible through the ADR.

There are many existing metadata standards and schemas that serve the needs of many different user communities. Current metadata standards and standards for crosswalks exist in the library and archival world. Dublin Core to MARC crosswalks, MARC to MODS, and Learning Object Metadata (LOM) to Dublin Core constitute only a few examples. Childress, Godby, and Young (2004) discuss building a repository of effective

metadata crosswalks for large databases in order to ensure consistency and standardization between schemas. Mappings of the metadata must be endorsed by a stakeholder community and relationships between source and target metadata standards must also be well-defined. A repository of metadata crosswalks and standards for the library community provides a wonderful step in the direction of interoperability between schemas and metadata and would fill a need, especially in the digital library community.

Non-library schemas such as Stat Crew, however, present a dilemma since crosswalks must be built with no standardization. Interoperability must also be considered. These issues make the conversion and crosswalking of metadata and schemas much more difficult.

In the case of our project, our main concern, once appraisal is completed, is building the crosswalk and converting it with an XSL. Several questions remain open and unanswered even beyond the available funding and the degree of time allotted to the project. Some questions that must be considered include:

1. How much loss of data between the two schemas should be allowed? What is the amount of granularity that we hope to capture? In a perfect environment, each schema element and attribute from the source schema could be paired or matched with similar elements and attributes in the target standard and transformed. Such exact one-to-one pairing or synchronizing is rare, however, and properties may not precisely crosswalk or map. This factor remains a consideration since the content of both schemas will not always schematically match. In other words, when elements or attributes mismatch, they will not translate well. Therefore, we will need to adapt and may necessarily lose or forfeit some granularity of data due to element source and target mismatch. One or both of our schemas may contain extra elements and descriptors that cannot be paired with the other schema. Will the Athletics Department be willing to depart with, or even lose, some data or granularity of data?
2. If funding allows, will the DLT need to hire a librarian who has the technical skills to build and develop crosswalks and XSLTs, or will the team hire a librarian and a separate IT specialist conversant in those areas? Which would be less expensive and/or more efficient? Can a librarian with these technical skills be found?
3. Finally, if the crosswalk cannot be built, due to funding or technical issues, what is the possible alternative? At minimum, how do we want to preserve the Stat Crew Data? How and where will it be preserved?

In the library world, crosswalks and metadata mapping standards are becoming more and more the norm. Outside of this world, however, translating metadata and schemas becomes more challenging. Businesses and industries must increasingly share their data and information with each other. Metadata standards are in place, or taking shape, and interoperability has emerged as a key concept. Yet even as this process progresses, the

expertise to develop crosswalks seems problematic. Many metadata schemas and standards are developed independently. Others may be proprietary. Some use different methods, processes and terminology. The challenge for archivists and librarians is to take metadata, metadata standards and schema that have been created for narrow or proprietary purposes and make them understandable, as well as preserve them over the long term. This must be done without diluting them to the point where they lose their meaning. In the case of Stat Crew and SportsML we are finding these issues very challenging. Since we are at the forefront of this research, our experiences may show others that two disparate schemas can be preserved, transformed and viewed by a wide audience.

Definitions

Crosswalk — The relationships between the elements of two or more data structures. 2. A chart or diagram that indicates the correspondence between two systems.

Document Type Definition (DTD) — A set of rules that specify the structure of a document and the tags used to define that structure and that can be used to validate whether a document is well formed.

Extensible Markup Language (XML) — A standard to promote sharing information over the Internet by specifying ways to describe the information's semantic structure and to validate that the structure is well formed.

Extensible Stylesheet Language Transformations (XSLT) — A metalanguage that describes rules to convert an XML document with one set of tags to a different XML document with different tags, effectively changing the structure of the document.

Interoperability — The ability of different systems to use and exchange information through a shared format.

Schema — A formal description of a data structure.

Text Encoding Initiative (TEI) — Guidelines for encoding in Standard Generalized Markup Language (SGML) or Extensible Markup Language (XML) machine-readable texts of interest to the humanities and social sciences.

Pearce-Moses, Richard. [A Glossary of Archival and Records Terminology](#). Society of American Archivists (2005).

Reference

Childress, E., Godby, C.J., & Young, J. (2004). *A Repository of Metadata Crosswalks*. *D-Lib Magazine* 10 (12).

DMReview.com Glossary: <http://www.dmreview.com/glossary/n.html>

Dublin Core Metadata Glossary, Final Draft, Feb. 24, 2001, Online. Available at <http://library.csun.edu/mwoodley/dublincoreglossary.html>

Fedora: <http://www.fedora.info/>

IBM Glossary of Unicode Terms:

<http://www.ibm.com/developerworks/library/glossaries/unicode.html>

Normalization Theory for XML:

[http://64.233.167.104/search?q=cache:KRdrvO5V8csJ:www.sigmod.org/sigmod/record/issues/0612/p57-column-](http://64.233.167.104/search?q=cache:KRdrvO5V8csJ:www.sigmod.org/sigmod/record/issues/0612/p57-column-libkin.pdf+normalization+xml&hl=en&ct=clnk&cd=5&gl=us&client=firefox-a)

[libkin.pdf+normalization+xml&hl=en&ct=clnk&cd=5&gl=us&client=firefox-a](http://64.233.167.104/search?q=cache:KRdrvO5V8csJ:www.sigmod.org/sigmod/record/issues/0612/p57-column-libkin.pdf+normalization+xml&hl=en&ct=clnk&cd=5&gl=us&client=firefox-a)

<soaprpc/>: <http://www.soaprpc.com/glossary.html>

Sports Markup Language: <http://www.sportsml.org/index.php>

Stat Crew Software: <http://www.statcrew.com/>

Does your university archives have born-digital records?

Share how you are effectively managing these digital records by submitting a case study to Campus Case Studies.

Visit www.archivists.org/publications/epubs/CampusCaseStudies/.