Indexing into Controlled Vocabularies with XML

Stephen C. Arnold
Bradley J. Spenla
College of Information Science and Technology
Drexel University
Philadelphia, PA, 19104, USA
stephen.arnold@cis.drexel.edu
brad.spenla@cis.drexel.edu

Abstract
This paper presents an architecture to place index references from control vocabularies into XML documents. This provides three advantages currently not present in XML: the ability to simplify the writing of complex document type definitions that refer to control vocabularies, the ability to write XML documents with better indexing, and better retrieval of XML documents.

1. Motivation

One of the major objectives of XML is to allow documents to be easily retrieved. [5] To this end,

- XML is a semantic markup language,
- XML arranges the contents of documents into hierarchical structures, and
- XML provides a method of sharing document type definitions [1].

All of the above will make retrieving documents easier, but a major approach used for making document retrieval easier is not included in XML. That approach is indexing documents using control vocabularies.

A document that has been indexed using a controlled vocabulary is much easier to find than one that has not [10]. The process of searching for a document eventually becomes an issue of matching words. The individuals looking for the document specify some set of words or phrases (terms) they wish to find. They hope the documents returned from a search will be closely and significantly related to the set of terms that they provided.

Because natural language is highly ambiguous, simply matching the exact terms given in a search often results in a set of documents that are not closely or significantly related. There are two fundamental problems. One, polysemy, many of the documents retrieved may use the terms that were specified in the search in a manner different from the way that the searchers intended. Two, synonymy, many documents may have been excluded because the documents do not contain the terms specified in the search, even though they do contain some term that has the same meaning [10].

Library and Information Science has seriously addressed this issue of finding documents through the development of indexing using control vocabularies. Controlled vocabularies normally have the characteristic that terms have one and only one definition [14]. By associating (indexing) documents or sections of documents with terms from a controlled vocabulary, the two problems described above are alleviated. Problem one is resolved because the terms used in indexing are allowed to have only one meaning. Even if the document uses the exact word or phrase in the term, the indexer will (should) determine that it is a different meaning and not associate the document with this term from the controlled vocabulary. A document may use a word or phrase that is different from a term in the controlled vocabulary but has the same meaning as some term in the controlled vocabulary. Then the indexer will (should) index the document with that term. This alleviates problem two by insuring that documents are indexed by terms even if the term does not exist in the document.

Indexing documents from controlled vocabularies makes searching for documents even more effective because of two additional aspects of controlled vocabularies: synonym lists and the relationships between terms that are part of a controlled vocabulary [11]. A synonym list is a list of words or phrases that often are used in documents to mean the same thing as a term in the controlled vocabulary. These lists are of great use to both indexers and searchers. Indexers can use these lists to find a word or phrase in the document that means the same thing as a term in the controlled vocabulary. Then the indexer can index the document with the term from the
controlled vocabulary, even though the term did not exist in the document.

Searchers often do not know all the terms in a controlled vocabulary. They need to find the term that most closely matches the concept for which they are searching. They can attempt to match the words or phrases they associate with the concept they wish to retrieve. After finding the term, they can use it in the search. Searching a controlled vocabulary for the right term, and then searching a set of documents indexed against the controlled vocabulary is a much easier task than searching a large document store without indexing using a controlled vocabulary [2].

The architecture to integrate controlled vocabularies into XML consists of two parts: a controlled vocabulary reference description that can be inserted into an XML DTD and a set of software tools that interpret and expand this reference. The controlled vocabulary reference is not intended to be a general-purpose extension of the XML language but a special-purpose extension for highly indexing documents. Specifically, it is a new clause added to XML document type definitions (DTDs) that defines a replacement of a clause or rules with terms from a controlled vocabulary. After the reference is processed, the result is a standard DTD containing the expanded CV references. The set of software tools provides functionality to users of these tools that takes advantage of these CV references.

2. Architecture

The architecture to integrate controlled vocabularies

2.1 IDTDs

Indexing document type definitions (IDTDs) are XML DTDs with an indexing expression. This expression consists of two parts: controlled vocabulary reference and usage. The controlled vocabulary reference is a specification of some set of terms from a controlled vocabulary. The usage describes how these terms are to be used in the DTD.

The controlled vocabulary reference (CVR) is an expression that evaluates to a set of terms from the controlled vocabulary. This expression consists of a set of simple CVRs combined by set operators (union, intersection, set subtraction). In the controlled vocabulary, the authors are assuming each term has some unique label associated with it. A simple CVR contains a unique term or label of a term followed by some expression of relationships between this term and other terms in the controlled vocabulary. These relationships are either descriptions of the hierarchy or related terms in the controlled vocabulary’s structure.

For example, <$p <LeftHand> $s <0> $e <2>> means the set of descendents of the term “LeftHand”, where the descendents are a depth of zero to two from the root of the subtree in the controlled vocabulary starting at the node containing the term “LeftHand”. With this controlled vocabulary that includes: “A.2.3_leftHand”, “A.2.3.1_leftPinky”, “A.2.3.2_leftRingFinger”, “A.2.3.3_leftMiddleFinger”, “A.2.3.4_leftIndexFinger”, “A.2.3.5_leftThumb”.

Note that the full label (the label of a term and the labels of its ancestors concatenated together) is included in the returned result. The usefulness of this will become apparent in the Query Language Application section below.

The usage section of an indexing expression describes how the list of terms generated by indexing expression is
used in the DTD. There are two general ways that the terms are used: lists or replication. In a list usage, the terms are concatenated into a list and this list is substituted for the CVR in translating the IDTD into a standard DTD. To make a list from the terms, the delimiter between list items also needs to be specified. The replication makes one copy of the rule in which the indexing expression exists for each term produced by the CVR. The terms are substituted for the indexing expression in translating the IDTD to a standard DTD. In addition, information on how to translate the terms into strings that will be able to be used in a standard DTD needs to be specified such as the characters used for quoting the terms. All this information exists in the usage specification. A full example of both the list and replication usage are given in the translator section below.

2.2 Tools

The set of software tools are programs to support the following tasks: using IDTDs with XML tools that are not built to take advantage of the IDTD indexing expressions, writing XML documents that have references into controlled vocabularies, and querying XML document bases using controlled vocabulary references.

The major issue in indexing from controlled vocabularies in DTDs is that the indexing reference is a single clause while its represents multiple terms form the controlled vocabulary. The number and value of the terms exists external to the IDTD. It exists in the CV. The IDTD only specifies the method of expansion for the reference. This means the CV can change independent of the reference so long as the base term of the reference still exists within the CV.

The logic for converting an IDTD to a DTD is more complex than a simple macro expand. In some cases the terms referred to by a CV references generate a list. That list may represent a list of items required in the body of the XML document defined the DTD. That list may also be one choice for a word within the body of the XML document. Lastly, this list can be used to specify a series of DTD rules.

2.2.1 Translator. A translator takes an IDTD and generates a DTD. This system is useful in two ways. Translating from IDTDs to DTDs makes it possible to use all the existing XML software that uses standard DTDs. Translating from IDTDs to DTDs allows the writers of DTDs to write what would be very large DTD specifications in a very few lines.

A large number of software tools have been written that use DTDs. There are XML parsers [2] [9], editors to assist in writing XML documents [8] [17], and Web servers that display XML documents (using XSL specifications) [1]. IDTDs are an XML extension that does not work with theses software systems. Without a translator, IDTDs would be of little use because of lack of integration with existing systems.

There are many large controlled vocabularies such as ICD-9 codes and the Dewey Decimal Classification. A standard DTD that uses a signification section of these controlled vocabularies would be very large. An IDTD can be much smaller. The indexing expressions also represent a large list of terms at a more symbolic level. Smaller and more symbolic source code is easier to maintain [4]. Also, controlled vocabularies change over time. New diseases are added to the ICD-9 codes and new topic areas are added to the Dewey Decimal Classification. If a standard DTD enumerated the terms out of these controlled vocabularies, the DTDs would need to be rewritten each time the vocabularies change. With IDTDs, only the translator needs to be rerun to integrate the changes in the controlled vocabulary.

The authors have constructed a prototype translator. Below are two examples that use a simple controlled vocabulary to build a DTD defining an injury report and a body assessment.

The injury report records who was injured, who is the attending doctor and what part of the body was hurt. It uses an indexing expression to describe the body part that was injured.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT INJURY_REPORT (PATIENT, AREA_AFFECTED+, ATTENDING)>
<!ELEMENT PATIENT (FNAME, LNAME)>
<!ELEMENT FNAME (#PCDATA)>
<!ELEMENT LNAME (#PCDATA)>
<!ELEMENT AREA_AFFECTED EMPTY>
<!ATTLIST AREA_AFFECTED id (
| > | $l <A.2> $s <1>
$e <4>>
) #REQUIRED>
```
Figure 1. IDTD for an injury report.

Note that the author of the IDTD did not have to specify each body part, only a reference to the human body controlled vocabulary. Otherwise the author of the IDTD would need to write.

<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT INJURY_REPORT (PATIENT, AREA_AFFECTED+, ATTENDING)>  
<!ELEMENT PATIENT (FNAME, LNAME)>  
<!ELEMENT FNAME (#PCDATA)>  
<!ELEMENT LNAME (#PCDATA)>  
<!ELEMENT AREA_AFFECTED EMPTY>  
<!ATTLIST AREA_AFFECTED id (A.2.1_leftUpperArm|A.2.2_leftForearm|A.2.3_leftHand|A.3.1_rightUpperArm|A.3.2_rightForearm|A.3.3_rightHand|A.3.1_leftThigh|A.3.2_leftCalf|A.3.3_leftFoot|A.4.1_rightThigh|A.4.2_rightCalf|A.4.3_rightFoot) #REQUIRED>  
<!ELEMENT ATTENDING (FNAME, LNAME)>  

Figure 2. DTD derived from IDTD for an injury report.

The body assessment DTD is intended to guide a person through a complete assessment of a person’s body. Here the controlled indexing expression is used not to specify attribute values, but elements.

<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT ASSESSMENT_REPORT ANY>  
<!ELEMENT A_body (#PCDATA)>  
<!ELEMENT A.1_head (#PCDATA)>  
A.2.1_leftArm (#PCDATA)>  
33 additional rules  
A.3.1_leftThigh|A.3.2_leftCalf|A.4.3_rightFoot) #REQUIRED>  
<!ELEMENT ATTENDING (FNAME, LNAME)>  

Figure 4. DTD derived from IDTD for a body assessment.

2.2.2 Editor. Indexing documents is not a simple task. As with desktop publishing tools [13], XML authoring tools are going to need functionality to support indexing. Editors have already been produced to help in writing the content of XML documents. These tools are structured editors that build the required structure of an XML document as specified by the DTD associated with the document. These help the authors of the XML documents by only allowing them to write valid XML documents but provide no support for indexing using controlled vocabularies. The authors are prototyping an editor to provide support for indexing using controlled vocabularies.

IDTDs specify XML documents with indexing expressions. Because controlled vocabularies are structured into trees, the standard techniques for displaying trees on computers can be used to display the potential values for an indexing expression. The additional information of terms definitions, related terms, and synonym lists can also be displayed to assist the author of the document in picking the correct term for an indexing expression.

The authors are building a prototype of XML editor that provides support for indexing expressions by presenting the tree structure of a controlled vocabulary. When the indexing expression substitutes for an attribute value and is highlighted by the author of the XML document, the editor presents the author with a tree display of the terms allowed to be substituted for the expression. In addition, if the author highlights a term from the controlled vocabulary, the editor displays the additional information about the term, such as its definition. Below is a screen design the authors are considering for this system.
Figure 5. Editor for writing XML documents defined by IDTDs.

2.2.3 Query Language Application. CV references in an IDTD can be used as a query language application. When a term is placed into the body of an XML document, its full label is also placed in the document at the same location. The label is a unique reference into the structure of a controlled vocabulary. The label is the data needed to determine where a term is within this structure. The structure can be used to support the search. The example of finding documents containing the term “thumb” by searching for “arm” and its descendents is given in the Motivation Section.

The rest of the discussion about query language extension is discussed in the Future Work section because the authors have not yet made any significant progress in constructing these yet.

3. Future work

3.1 Indexing Extension

The authors need to examine what kinds of sets of terms are useful and meaningful to be able to extract from a controlled vocabulary for this application and how the people who will be writing IDTDs most easily and effectively express this. This may be done by examining existing indexed document bases or by studying the people who are expected to author IDTDs.

3.2 Tools

3.2.1 Translator. The usage section of an indexing expression is currently a simple macro replace. It is quite easy to write IDTDs that translate to syntactically invalid DTDs. Much of the information from the usage section should be able to be determined from the syntax of the IDTD. The translator will also be extended to work with multiple controlled vocabularies in the same IDTD. The authors intend to refine the translator to be able to do these tasks.

3.2.2 Editor. The editor exists only as the beginning of a prototype. The prototype needs to be completed, the editor needs to be built and studied. The authors intend to study how well it helps authors index the documents appropriately. Alternate designs will also be considered to improve the usability of the editor.

3.2.3 Query Language Application. Querying based on controlled vocabularies is a well-understood process. The structure of the controlled vocabulary can be used directly within a query. An example of how the structure can be effectively used to produce a more meaningful result set is given in the Motivation Section.

The XML query language [17] is currently being defined. When its definition is more stable, the authors will look at how to extend it to effectively query against a document base of documents produced from IDTDs. This extension will enable queries to contain references to controlled vocabularies and take advantage of the structure contained in the controlled vocabulary to produce more meaningful result sets.

These queries will be able to look for index terms and multiple combinations of index terms based on relationships between the terms in the controlled vocabularies. It may be able to use the synonym list to attempt to find documents that should be indexed by a term that uses that terms synonyms.

The controlled vocabularies may also be useful in helping people construct queries. A tool that presents the section of the controlled vocabulary referred to in a DTD while a person is constructing a query may be useful in helping the person to pick the correct term to retrieve the documents desired.

3.3 XML Schema

Like much of XML, XML Schema is still being defined. [16] It is a possible replacement or alternate for XML DTDs. It, like DTDs, does not have a mechanism for indexing using a controlled vocabulary. It may be possible to introduce a new data type for controlled vocabularies, but this would not provide the ability to define element names from a controlled vocabulary.

As XML schema becomes more stable, the authors will look at what is needed to effectively define an XML schema that includes indexing from controlled vocabularies.
3.4 Web Component Libraries

As web site and application development moves towards using libraries of components [7], the ability to easily find and reference these components will become critical. Controlled vocabularies will be one effective method of organizing these component libraries. XML is being used as control documents for managing component libraries for web applications. The authors will look into integrating the CV references into these XML-based component management systems.

3.5 XML Preprocessor Statement

The authors intend to design a second approach for including references into a controlled vocabulary that used the XML preprocessor statement and then compare the two designs. The current design has the advantage that all the semantics of the CV reference are at the same location in the file of the CVS reference. The second approach has the advantage that it will not require special processing for the software to interpret the IDTD (but it will lose the meaning of the CV reference). The authors plan to determine which is more important to the targeted users: location of semantics or standardization of documents the users write.

3.6 Dublin Core

The current language used for defining CV references only reflects the tree structure of controlled vocabularies. The Dublin Core [6] defines a richer model for the relationship between entries within a controlled vocabulary. The authors will extend the language for defining CV references to express the concepts within Dublin Core.

4. Conclusion

Indexing from controlled vocabularies makes documents easier to find. One of the objectives of XML is to make documents easier to find. XML does not have a mechanism to support indexing from controlled vocabularies. IDTDs are a special purpose addition to XML DTDs that integrates indexing using controlled vocabularies. This addition and the tools for supporting it: a translator, an editor, and a query language extension will enable XML documents to be able to be indexed using controlled vocabularies effectively making them even easier to find.

5. References


