Conventions:

Proposed XML Format for Glossaries

A recommendation for transporting dictionary information (GlossXML)

Author: Dominic John *Repici*

- Contents
  - Overview
  - Design Philosophy & Goals
  - Design Motivation
  - XML Tags Used
  - Tag Attributes
  - **Using Edit Communities**
    - Edit Communities & Context
  - Aggregating Glossaries
  - SubEditor Elements
  - The DTD
  - Permissions
  - Related Standards

Overview

This paper presents a recommendation for "GlossXML", an XML file format to be used to transport dictionaries of words and phrases. The XML is presented first as pros, then as a DTD. This paper is currently a work in progress. Changes to DTDs and schemas may lag behind the other presented information while work is moving forward.

Design Philosophy & Goals

In a word, the design philosophy for this project is ***Simplicity***. The specification defined here shall be easy to understand, easy to implement, and easy to use. It must also be concisely and unambiguously defined, flexible, and enabling of applications beyond those that can be foreseen at the time of initial definition.

Lastly, while this specification will stand on its own it will be designed in such a way that it shall be capable of serving as the basis for an RPC-style transactional convention to

be developed in the future.

Design Motivation

While there are many other standards for transporting dictionary information, those researched were found to be unsuitable for this specific project for one reason or another. Though many of the existing standards are excellent and some very ambitious, none were found to be a good fit for the functionality required by this project.

In the case of the more ambitious standards, such as MARTIF for example, concept-based categorization was the primary design goal and well represented in the standard. The primary design goal for this project however, is interoperability. Though, it should be mentioned, a new push is currently underway for something called "blind" MARTIF which would support interoperability with arbitrary external implementations of itself.

A more important consideration for this project, and one no existing standards meet, is that of recognizing communities as a fundamental structural attribute of language and meaning. In my opinion, languages are rooted in human communities, and communities are the top-most context in the meanings of phrases used by individuals. An example of this is described for the term *hard drive* in the **Communities and Context** section below.

I must stress here, this is only my opinion and I'm not a language expert. I'm in no way arguing that existing standards which don't include a notion of communities as top-level context are deficient. In fact, it is much more likely those who have produced these other standards have some insight or know something about the problem domain that I do not, and have excluded communities as contextual attributes deliberately.

XML Tags Used

- **<Lexicon>**
  The document element of the export file. This may contain multiple <Entry> elements. One (or possibly more) for each term in the glossary.
  - **<Entry>**
    A word or phrase in the glossary. Will contain a sub-set from the set of elements defined below.
    Zero or more Entry elements in each Lexicon.
    - **<Term>**
      The term, word, or phrase defined by this entry.
      Contains #PCData. **One** in each entry.
    - **<Definition>**
      The text of the definition for the term. May include CUF™ format codes.
      Contains #PCData. **Zero or more** in each entry.
    - **<SeeAlso>**
      Other terms that are related to the Term.
      Contains #PCData. **Zero or more** in each entry.
    - **<Synonym>**
      One synonym term for the Term. May be classified with a part of

speech attribute or left un-classified.
Contains #PCData. **Zero or more** in each entry.

- **<Abbreviates>**
  A special-case synonym used only when the primary term is an abbreviation or acronym for another term. This element may contain the expanded (unabbreviated) form of the term. Expansion must only be performed once on the term. That is, if the expanded form of the term contains another abbreviation or acronym, the contained abbreviations will remain intact.
  Contains #PCData. **Zero or more** in each entry.
- **<Antonym>**
  An element containing a term that is the antonym of the entry in the <Term> element.
  Contains #PCData. **Zero or more** in each entry.

- **<Pronunciation>**
  One pronunciation guide for the Term.
  Contains #PCData. **Zero or more** in each entry.
- **<Translation language=*lang*>**
  Term translation to a different specified language. The element contains the term translated to the language specified in the *language* attribute. Languages must be specified using the two or three character representations documented in ISO 639-1/2 and may optionally contain a two character language-code modifier (e.g. eng-**us**)
  Contains #PCData. **Zero or more** in each entry.
- **<Etymology>**
  Element holding notes on term's etymology.
  Contains #PCData. **Zero or one** in each entry.
- **<UsageNotes>**
  Element holding notes on term's usage.
  Contains #PCData. **Zero or more** in each entry.
- **<UsageCitation>**
  Element holding examples of term in use.
  Contains #PCData. **Zero or more** in each entry.

- **<EditGlossarist>**
  The glossarist who last edited or created the entry. Can be a name, code, IP-address, etc.
  Contains #PCData. **Zero or one** in each entry.
- **<EditDate>**
  The date this entry was last edited.
  Contains #PCData. **Zero or one** in each entry.
- **<EditCommunity>**
  The Name or title of the community where this term was last edited or produced.
  Contains #PCData. **Zero or one** in each entry.
- **<EditCommunityURL>**
  The URL of the community where this term was last edited or produced. Should be a valid URL reachable on the Web.
  Contains #PCData. **Zero or one** in each entry.

- **<SubEditor>**
  The Name or title of the compilation where this term had minor edits made.
  Contains #PCData. **Zero or one** in each entry.
- **<SubEditorURL>**
  The URL of the compilation where this term had minor edits made. Should be a valid URL reachable on the Web.
  Contains #PCData. **Zero or one** in each entry.

[top]

Tag Attributes

Some tags may have attributes. All attributes are optional unless otherwise noted.

- **<Lexicon>**
  - **Language=*"language"***
    The language the lexicon is written for if any specified. Languages must be specified using the two or three character representations documented in ISO 639-1/2 and may optionally contain a two character language-code modifier (e.g. eng-**us**)
  - **Subject=*"subject label"***
    The special field or discipline the lexicon is limited to if any. Examples of subject labels are: 'Law', 'Math', 'Naut', but a short descriptive title is also acceptable here.
  - **EditCommunity=*"CommunityTitle"***
    The name of the community responsible for this dictionary.
  - **EditCommunityURL=*"CommunityURL"***
    The URL for the community named in EditCommunity.
- **<Entry>**
  - **Language=*"lang"***
    An optional attribute naming the language that the primary term/definition pertains to. SHOULD only be used in mixed language applications, such as translators, where the <Entry> language is not equal to the language specified in the <Lexicon> element. Languages must be specified using the two or three character representations documented in ISO 639-1/2 and may optionally contain a two character language-code modifier (e.g. eng-**us**)
  - **Subject=*"subject label"***
    An optional attribute naming the subject that the primary term/definition pertains to. SHOULD only be used in mixed subject lexicons, or when the Entry subject differs from the subject specified in the Lexicon element.
- **<Definition>**
  - **PartOfSpeech=*"Parameter"***
    Parameter may be one of:
    1. Noun
    2. Verb
    3. Adjective
    4. Adverb

5. Idiom
6. Unclassified (Note: Same as no attribute)

- **<Synonym>**
  - **PartOfSpeech=*"Parameter"***
    Parameter may be one of:
    1. Noun
    2. Verb
    3. Adjective
    4. Adverb
    5. Idiom
    6. Unclassified (Note: Same as no attribute)
  - **Sense=*"Parameter"***
    Parameter may be one of:
    1. Concise
    2. Broad
    3. Idiomatic
    4. Unclassified (Note: Same as no attribute)
- **<Antonym>**
  - **PartOfSpeech=*"Parameter"***
    Parameter may be one of:
    1. Noun
    2. Verb
    3. Adjective
    4. Adverb
    5. Idiom
    6. Unclassified (Note: Same as no attribute)
  - **Sense=*"Parameter"***
    Parameter may be one of:
    1. Concise
    2. Broad
    3. Idiomatic
    4. Unclassified (Note: Same as no attribute)
- **<Translation language=*lang*>**
  - **Language=*"lang"***
    A **MANDATORY** attribute naming the language that the term is translated to within the element. Languages must be specified using the two or three character representations documented in ISO 639-1/2 and may optionally contain a two character language-code modifier (e.g. eng-**us**)
  - **Quality=*"Parmameter"***
    An optional attribute denoting the class of the translated term (not of the primary term)
    Parameter may be one of:
    1. Literal
    2. Concise
    3. Loose
    4. Sense
    5. Unclassified (Note: Same as no attribute)
  - **PartOfSpeech=*"Parameter"***
    An optional attribute denoting the part of speech of the primary term represented in this translation.
    Parameter may be one of:

1. Noun
2. Verb
3. Adjective
4. Adverb
5. Idiom
6. Unclassified (Note: Same as no attribute)

Using Edit Communities

EditCommunity elements ("**<EditCommunity>**" and "<**EditCommunityURL>**") allow for tagging definitions with the groups responsible for producing them. Asside from giving content providers an easy way to distribute copyright credits with their definitions, these are also useful for imparting a community context to the meaning. They may also be used to rate the quality of a particular definition as it pertains to a given discipline. For example, such ratings may be done:

- **Manually:** A viewer may take note of the community (if displayed) and use it to make a personal assessment of the value of the definition, or,
- **Automatically:** Viewers from different fields may be asked to rate each definition based on the three-Cs ("Correctness", "Conciseness", and "Clarity"). These ratings can then be automatically compiled and analyzed in a variety of ways to determine the performance of the originating community's definition within a given context.

Lessons from the Tower
Communities & Context

Community ratings are not limited to acting as one-dimensional qualitative indicators, they are multiplied by **context**.

For example, a viewer looking up the term ***"Hard Drive"*** may consider a definition from a hardware engineering community (e.g. IEEE) and a definition from a community of software developers (e.g. ACM) and choose the one most appropriate to the **context** where s/he is using it.

Here, the definition from the hardware community will likely describe "hard drive" as a piece of electronic equipment, while the definition from the software community will define the term as it applies to software applications and information storage.

This should not discount the historical qualitative aspects of community ratings however. A phrase from say, "Mike's Electronic Enthusiast Page" may or may not be judged by the viewer to be the better choice, depending on past experience with definitions from that community.

Aggregating Glossaries

For aggregation purposes, the ***EditCommunity*** elements are defined to default up through the following hierarchy:

3. The Web location and URL from where the entry was retrieved
    2. The EditCommunity **attributes** from the <Lexicon> tag
        1. The EditCommunity **tags** from within the <Entry> itself.

When aggregating glossaries, the aggregating system must attempt to *provide* the **EditCommunity** and **EditCommunityURL** tags within each Entry element. These tags should contain an attribution to the community responsible for originating the definition (such attribution will usually be required by the community via copyright restrictions). To accomplish this, when reading each <Entry> element the aggregating system must:

1. First look for EditCommunity *tags* in the <Entry> element itself and use those if they are there.
2. If no EditCommunity tags are in the <Entry> element, the system must then look for the EditCommunity *attributes* in the <Lexicon> element and assign those to the aggregated <Entry>.
3. If there are no EditCommunity attributes within the <Lexicon> tag then the aggregator must use the URL and (if available) title from the Web location where it is gleaning the dictionary for aggregation. It should assign each of these to the these <EditCommunityURL> and <EditCommunity> tags for the element respectively.

Aggregating systems may store community information in whatever form they wish but MUST provide it in such a way that it is exactly the same as it would be if they had stored the edit community and URL with each individual term when gleaning the information in accordance with the steps defined here.

SubEditor Elements

Subeditor elements ("**SubEditor**" and "**SubEditorURL**") are elements of an <Entry>. They allow organizations to make minor edits, such as spelling, and punctuation, to definitions produced by outside communities while preserving the copyright connection to the community who originally produced the definition.

These must ONLY be added to a given **<Entry>** if edits are made to the entry. Those outside of the originating community making only minor changes to entries must not alter the EditCommunity... tags of the entry, but should instead claim responsibility using SubEdit elements.

This functionality can be especially useful to organizations who want to specialize in collecting ratings and compiling glossaries of the best definitions from other sources. It allows such organizations to provide clean- up editing to some entries while still maintaining copyright requirements.

The DTD File

In keeping with the design goal of a precise and unambiguously defined convention, a DTD is included as part of this recommendation.

The document element for this format is <Lexicon>. An XML file may be made valid by including a reference to the DTD file in a <!DOCTYPE ...> element, or may simply follow

the rules provided in order to be a well formed document of this type.

Here's an example of how a valid XML file in this format might reference an external DTD file.

```
<!DOCTYPE Lexicon SYSTEM "GlossXML.dtd">
```

GlossXML.dtd

```
<!--
= The following DTD is:
=
=     (C) Copyright 2002, Creativyst, Inc.
=           ALL RIGHTS RESERVED
=
= For more information go to:
=         http://www.Creativyst.com
= or email:
=         Support@Creativyst.com
=
=
= You may use and distribute this XML DTD
= freely as long as this entire notice remains
= with it and intact and the entire DTD is distributed
= UNMODIFIED from its original content and format.
=
=
-->


  <!ELEMENT Lexicon (Entry*)>

  <!ATTLIST Lexicon Language CDATA #IMPLIED
                    Subject CDATA #IMPLIED
                    EditCommunity CDATA #IMPLIED
                    EditCommunityURL CDATA #IMPLIED>



  <!ELEMENT Entry
             (Term,
             (Definition*       |
              SeeAlso*          |
              Synonym*          |
              Abbreviates*      |
              Antonym*          |
              Pronunciation*    |
              Translation*      |
              Etymology?        |
              UsageNotes*       |
              UsageCitation*    |
              EditGlossarist?   |
              EditDate?         |
              EditCommunity?    |
              EditCommunityURL? |
```

```
              SubEditor?        |
              SubEditorURL?        )*)>

   <!ATTLIST Entry Language CDATA #IMPLIED
                   Subject CDATA #IMPLIED>


   <!ELEMENT Term (#PCDATA)>
   <!ELEMENT Definition (#PCDATA)>
   <!ELEMENT SeeAlso (#PCDATA)>
   <!ELEMENT Synonym (#PCDATA)>
   <!ELEMENT Abbreviates (#PCDATA)>
   <!ELEMENT Antonym (#PCDATA)>

   <!ELEMENT Pronunciation (#PCDATA)>
   <!ELEMENT Translation (#PCDATA)>
   <!ELEMENT Etymology (#PCDATA)>
   <!ELEMENT UsageNotes (#PCDATA)>
   <!ELEMENT UsageCitation (#PCDATA)>

   <!ELEMENT EditGlossarist (#PCDATA)>
   <!ELEMENT EditDate (#PCDATA)>
   <!ELEMENT EditCommunity (#PCDATA)>
   <!ELEMENT EditCommunityURL (#PCDATA)>
   <!ELEMENT SubEditor (#PCDATA)>
   <!ELEMENT SubEditorURL (#PCDATA)>

   <!ATTLIST Definition PartOfSpeech
                       (Noun        |
                        Verb        |
                        Adjective   |
                        Adverb      |
                        Idiom       |
                        Unclassified) "Unclassified">

   <!ATTLIST Synonym PartOfSpeech
                       (Noun        |
                        Verb        |
                        Adjective   |
                        Adverb      |
                        Idiom       |
                        Unclassified) "Unclassified">
   <!ATTLIST Synonym Sense
                       (Concise     |
                        Broad       |
                        Idiomatic   |
                        Unclassified) "Unclassified">


   <!ATTLIST Antonym PartOfSpeech
                       (Noun        |
                        Verb        |
                        Adjective   |
                        Adverb      |
                        Idiom       |
                        Unclassified) "Unclassified">
   <!ATTLIST Antonym Sense
```

```
                    (Concise     |
                     Broad       |
                     Idiomatic   |
                     Unclassified) "Unclassified">


<!ATTLIST Translation Language CDATA #REQUIRED>
<!ATTLIST Translation Quality
                    (Literal     |
                     Concise     |
                     Loose       |
                     Sense       |
                     Unclassified) "Unclassified">
<!ATTLIST Translation PartOfSpeech
                    (Noun        |
                     Verb        |
                     Adjective   |
                     Adverb      |
                     Idiom       |
                     Unclassified) "Unclassified">
```

Permissions

Permissions printed over DTD and schema files are supported as our permission statement for these constructs. Permissions to copy and distribute this recommendation unmodified and with credit and a link back to the Creativyst.com website are hereby granted. Permission to use this standard as the basis for information transport within products such as software applications is hereby granted provided it is used in unmodified form and credit and a link to the Creativyst.com website is included clearly within documentation. ALL OTHER RIGHTS, INCLUDING THE RIGHT TO MODIFY OR DERIVE OTHER WORKS ARE RESERVED.

Related Standards

There are currently many other proposed standards that encompass much of the functionality described in this paper. Some were researched prior to specifying this recommendation and will be listed here. Others will be added to this list as they are discovered over time.

Regarding those standards that were researched prior to this specification; in each case the existing standard or recommendation was deemed inappropriate for the narrow set of requirements defined within this specification. For example, existing standards are often designed to support language translations directly, where we only have a need to interface to these systems as a term glossary.

In some cases the existing standard was simply too broad or ambitious to be useable for this simple application, since concise interoperability across multiple independent implementations is required and may be diluted in broadly defined standards. In other

cases, the functionality may have been defined in platform specific language or targeted toward proprietary systems.

- **MARTIF**
  Machine-Readable Terminology Interchange Format (ISO 12200) - This ambitious and maturing standard is designed to handle the needs of ornthologists and language specialists. It is an offshoot of earlier work done by the TEI ('Text Encoding Initiative') and the Localisation Industry Standards Association (LISA). The 150 data categories used in MARTIF are standardized as ISO 12620. MARTIF is designed for concept-based systems. In their words: "[MARTIF] does not match the needs of non-concept-oriented approaches to terminology, i.e. lexicographic and NLP approaches, because MARTIF presupposes a *concept* orientation rather than a *word* orientation".
- **CLS Framework**
  The CLS Framework is the result of a joint effort of the Brigham Young University Translation Research Group (BYU TRG) and the Kent State University Institute for Applied Linguistics (KSU IAL). The framework deals with the structure and content of terminological databases (called "termbases"). The Framework can be used for representation of existing termbases, design of new termbases, and for the sharing of terminological data. Because CLS borrows heavily from MARTIF, many MARTIF resources can be found at this site.
- **World Wide Lexicon**
  SOAP based system defined in Visual Basic ("dot-net") semantics.
- **SALT**
  A very ambitious recommendation that includes MARTIF. In this case SALT stands for "Standards-based Access to Lexicographical & Terminalogical multilingual resources." This standard is promoted as a "set of standards" to unite many lexagraphical and terminalogical dictionary functions. In their words:
  - *The service offered by these tools will provide access to, and re- use of, heterogeneous multilingual resources derived from both NLP-lexicons and human-oriented terminology databases. Particular emphasis will be given to deriving, integrating, and interfacing ontologies and data structures in translation and localization environments.*
  - (**Please Note: Do not confuse this SALT standard** with ***Speech Application Language Tags*** which is a popular but completely different standard promoted by the salt forum for extending HTML tags to include speech interactions.)

- ***More to come...***