
Implementing a Government-wide Semantic Solution to Thesauri

Kenneth Sall

Ronald Reck

November 2005

Abstract

The traditional approach to federal glossaries and acronym lists has been to use HTML, Microsoft Word, Excel, or PDF. This paper describes how a strawman DTD developed into a XML Schema based upon ISO and ANSI standards for thesaurus concepts and terminology. This in turn became a stepping stone for a pilot implementation using RDF-based *SKOS (Simple Knowledge Organization System)*, an emerging specification from the W3C. The thesaurus approach developed by the authors enables term gatherers and authors to input terms and definitions via Excel, via a web form, or by using RDF editors. This flexibility accommodates authors with no XML skills as well as those who are XML adept. Semantic technology can be applied to the SKOS markup to extract associated terms, as well as terms that are broader or narrower in meaning than a given concept.

Table of Contents

1. Introduction	3
2. Problem	3
3. Design Goals	3
4. Candidate Requirements	4
5. Early Thesaurus Attempts	5
5.1. Basic Thesaurus Terminology	5
5.2. Initial DTD and Sample Instance	7
5.3. XSLT-Generated Search Links	8
6. Thesauri Standards and Specifications	9
6.1. ISO 2788 - Developing a Thesaurus	9
6.2. ISO 1087 - Vocabulary of Terminology	10
6.3. ISO 704 - Principles and Methods	11
6.4. ANSI/NISO Z39.19 - Construction, Format, and Management	11
6.5. Additional Specifications and Thesauri Sites	12
7. SKOS - W3C Working Draft	13
8. Our SKOS Element Subset and Extensions	16
9. Pilot Solution	18
9.1. Pilot Environment	18
9.2. Use Cases	18
9.3. Assumptions	18
9.4. Spreadsheet Format and Birds Example	19
9.5. Data Submission and Conversion to SKOS	19
9.5.1. Data Entry via a Web Form	22
9.5.2. Uploading CSV Data	23
9.5.3. Iteration Character	24
9.6. Bootstrapping URIs	25
9.7. Committing Data to Persistent Storage	26
9.8. Querying the Data Store	27
10. Next Steps for SKOSaurus	30
11. Conclusion	31
Acknowledgements	31
Bibliography	31

1. Introduction

One of the most widespread needs among US federal agencies is to create a glossary or thesaurus of terminology that pertains to an agency's mission or to a particular line of business. Communication among members of any community of interest (e.g., the Intelligence Community) can be significantly enhanced by ensuring that the terminology used is associated with the same semantics. Capturing acronyms and term definitions so that they may be distributed in a variety of formats makes the content widely available to civil servants, contractors and citizens. While the traditional approach to federal glossaries and acronym lists has been to use HTML, Microsoft Word, Excel, or PDF, much more can be gained by using XML technology in general, and semantic technology in particular.

This paper discusses ISO and ANSI standards for thesaurus concepts and terminology, which have informed the development of an RDF (Resource Description Framework) vocabulary called SKOS (Simple Knowledge Organization System), a W3C Working Draft. Although the SKOS-based thesaurus technique piloted by the authors¹ is independent of any particular agency's glossary format, our motivation was the potential use of this technology by the IC MWG (Intelligence Community Metadata Working Group), the CAF (Chief Architects Forum), the FEA (Federal Enterprise Architecture) DRM (Data Reference Model) Working Group, and the XMLCoP (XML Community of Practice). It is our belief that as more and more agencies and communities of interest adopt the SKOS-based approach to thesauri, the US federal government could create a flexible collection of identically structured thesauri developed by individual groups that could potentially be aggregated into a government-wide thesaurus. [See [IC MWG](#), [CAF](#), [DRM](#), and [XML-CoP](#).]

2. Problem

In various government meetings, we observed that different working groups, communities of practice, and agencies were using Microsoft Word and Excel to define glossary terms. However, each group used a different format: a bulleted list, a table, a spreadsheet, a wiki web page, etc. In some cases, only the term, definition, and source were recorded. In other cases, many additional properties were collected. While the specific format selected may have met the immediate needs of each group, it did not contribute to interoperability across groups. Since we observed that quite a few of the terms were related to enterprise architecture and technology, it seemed only natural that a format that was conducive to exchange, discovery, and harmonization of definitions would be far more useful in the long run for all agencies and working groups.

The SKOS-based thesaurus approach developed by the authors enables term gatherers and authors to input terms and definitions via Excel, via a web form, or by using RDF editors such as [Protégé](http://protege.stanford.edu/) [http://protege.stanford.edu/]. This flexibility accommodates authors with no XML skills as well as those who are XML adept. By storing the definitions in XML, an agency can write unique XSLT and XSL-FO stylesheets that transform the content into customized HTML pages or PDF. Other agencies and citizens can extract a subset of terms from the raw XML or create their own XSLT stylesheets. Furthermore, semantic technology can be applied to the SKOS markup to extract associated terms, as well as terms that are broader or narrower in meaning than a given concept.

3. Design Goals

After participating in several government working groups in which various glossary formats were discussed, one of the authors suggested the following design goals for a [more general XML-based solution](#) [http://kensall.com/gov/glossary/XML-Glossary-Strawman.ppt] in January 2005.

- Standards-Based - XML element names are based on international standards such as ISO 2788, ISO 704, and ISO 1087.

¹The authors can be reached at <Kenneth.B.Sall@saic.com> and <rreck@iama.rrecktek.com>.

- Flexible - The Glossary Schema, although initially a strawman to stimulate discussion, is fairly flexible with few required elements, many optional elements, and several repeatable elements.
- Provides a Framework - Since so few elements are required, terms can be added even before definitions are known. These terms act as placeholders that are fully supported by the XSD and XSLT.
- Specialized - Each term may have multiple definitions so that different agencies may use the same term with their own specialized meaning, where necessary.
- Collaborative - Since an XSLT stylesheet is used to sort the terms alphabetically, many individuals can work on their own glossary fragments (XML instances of the Glossary XSD). At any time, the various contributions can be easily merged without manual editing.
- Leverages Links - Search links are automatically generated for each term by means of the XSLT both to help kick-start and to augment the definition.
- Provides a Stepping Stone to the Semantic Web - The XSD is a XML subset of SKOS, which is an RDF vocabulary that complements OWL.

4. Candidate Requirements

Next the design goals became a [springboard](http://kensall.com/gov/glossary/XML-Glossary-Strawman-NG.ppt) [http://kensall.com/gov/glossary/XML-Glossary-Strawman-NG.ppt] for candidate requirements presented to a working group in February 2005.

1. The glossary / lexicon / thesaurus should use XML syntax with a schema (DTD, XML Schema, or RDF-S) for validation.
2. It should be applicable to any government agency.
3. The schema should be available to any civil servant or citizen
4. The schema should not be overly complex.
5. The schema should contain few required elements and many optional and/or repeatable elements.
6. It should be relatively easy to add new terms to the lexicon. Payware should not be necessary for authoring.
7. It should be relatively easy to combine terms authored by different individuals and different agencies, if desired.
8. The elements in the schema should be chosen with ISO standards in mind, to the degree that this does not overly complicate the schema.
9. It should be possible to create an XSLT stylesheet based upon the model to display an XML glossary instance document as HTML in modern browsers (IE, Firefox).
10. It is desirable that the XSLT generate additional search links not in the source.
11. Multiple definitions of the same term must be permitted, with either same or different context.
12. The entire approach should foster a clean separation of collaborative roles:
 - a. Developer of schema vs. developer of stylesheets.
 - b. Author/collector of terms and definitions.
 - c. Reviewer/approver of definitions.

-
- d. Consumer of results (e.g., agency with custom XSLT).
13. It should support semantic relationships between terms including related-to and synonyms.
 14. An approval process should be defined, but it should not interfere with contributions. Un-reviewed definitions would still be accessible, but without the stamp of approval.
 15. For each term, it must be possible to indicate:
 - a. Source (agency, author, document, and/or URL)
 - b. Context
 - c. Approval status
 16. Clear authoring conventions should be established
 - a. Case convention (UpperCamelCase, Title Case, lowercase, all caps?)
 - b. Pluralization (use singular form)
 - c. Compound terms (e.g., Data Architecture, Data Class)
 - d. Placement of acronym/abbreviation (separate element)
 - e. Placement of source/context/concept (separate element)
 - f. Citation method (URIs, bibliographical, free form?) Source could contain child elements for each possible format.
 17. Usage notes and/or examples are desirable.

5. Early Thesaurus Attempts

Our earliest investigation suggested that instead of simply developing a DTD or XML Schema for government glossaries, it would be more useful in the long run to create a schema that would accommodate a basic thesaurus. We realized that terms to be defined were likely to contain cross-references and relationships to other terms either within the same Community of Interest's vocabulary or to terms defined by another CoI. The [early work](http://kensall.com/gov/glossary/) [http://kensall.com/gov/glossary/] is located at: <http://kensall.com/gov/glossary/#older>

5.1. Basic Thesaurus Terminology

Before describing our early attempts, some background about thesauri may be helpful. [Our apologies to thesaurus experts if we over simplify the concepts discussed in this section. We were largely unfamiliar with most of this terminology at the outset of this project.]

A *thesaurus* is a list of concepts in a particular domain of knowledge together with explicit relationships. A *concept* is a unit of thought that exists in the mind as an abstract entity, independent of the term or terms that identify the concept. For each concept, related concepts are also included. A concept is represented by an *indexing term*, usually a noun or noun phrase. When a particular concept has more than one meaning, a *scope note* (SN) is often provided to restrict the domain, that is, to clarify or constrain the meaning. For example, the term "eagle" might mean a type of bird or a score two under par in golf. A scope note clarifies which of the two concepts we are indexing.

A concept may have one or more *related terms* (RT) which may be either synonyms or otherwise similar concepts. Hierarchical relationships are often expressed as well. A *broader term* (BT) is more general than the current concept

and is analogous to a parent node in a concept hierarchy (tree). A *narrower term* (NT) is more restrictive or specialized than the current concept; it is a child node in a concept hierarchy. For example, "bird" is broader than "eagle" and "bald eagle" is narrower than "eagle". Note that these relationships are reciprocal, so "eagle" is broader than "bald eagle" and "eagle" is narrower than "bird".²

Terms in a thesaurus are either *preferred* or *non-preferred*. A concept always has a preferred term (sometimes called the *descriptor*) which is term used consistently when indexing to represent the given concept. The preferred term is sometimes preceded by the keyword "USE" or "SEE" in a thesaurus. Often a concept also has one or more non-preferred terms, which are alternate labels for the concept. These are lead-in terms (entry points) such in an alphabetical index. Non-preferred terms are preceded by "USE FOR" (UF). For example, "Aves" is a non-preferred term for "birds", so a thesaurus might contain these entries:

- birds USE FOR Aves
- Aves USE birds (or: Aves SEE birds)

A *classification* is a grouping of related concepts as well as the separation of unrelated concepts such that the resulting groups are logical and in a useful sequence. Finally, a *concept scheme* is a set of concepts often including statements about semantic relationships between those concepts, such as a thesaurus or taxonomy.

Note that concepts in a thesaurus do not necessarily include definitions because semantics is conveyed by the various relationships that are identified. However, some thesauri do in fact include *definitions* (DEF), possibly labeled as scope notes (SN).

[Figure 1](#) depicts an excerpt from an unofficial version of the GAO (Government Accountability Office) Thesaurus from February 2005. The screenshot illustrates some of the terminology presented in this section. For additional terminology and more formal definitions, see [Glossary](#), [ISO-2788](#) and [Z39.19](#).

²The technical term for broader is *hypernym*; narrower is *hyponym*.

GAO Thesaurus February 2005	
<p>Legend: MT = Main term BT = Broader term RT = Related term NT = Narrower term SN = Scope note (a brief definition or description; how GAO defines the term)* UF = Used for (indicates non-preferred terms*) USE = GAO's preferred term</p> <p>* Example of UF and SN MT Academic achievement UF Scholastic achievement SN Student success in educational programs and activities. Compare Educational testing or Educational standards.</p> <p>(Example of USE) MT Academic institution accreditation USE Institution accreditation</p> <p>MT Administrative law BT Regulation NT Energy administrative law NT Environmental administrative law NT Federal administrative law NT Labor administrative law RT Agency proceedings RT Executive orders RT Freedom of information RT Judicial review RT Statutory law UF Administrative rulings UF Regulations (administrative) UF Rulings (administrative agencies) SN Rules and regulations issued by administrative and regulatory agencies to implement statutory law.</p>	<p>MT Adult education BT Education BT Employee training NT Continuing education RT Community colleges RT Compensatory education RT Higher education SN Educational programs/services, usually at the postsecondary level, for adults seeking personal, academic or occupational enhancement.</p> <p>MT Adults NT Elderly persons RT Parents RT Youth Status Accepted</p> <p>MT Advance appropriations BT Appropriations RT Advance funding RT Forward funding SN Provided by Congress for use in fiscal years beyond the year for which an appropriation act is passed.</p> <p>MT Advance funding BT Budget obligations RT Advance appropriations RT Forward funding SN Obligation and disbursement from succeeding year's appropriation.</p> <p>MT Advance payments BT Contractor payments BT Payments UF Prepayments</p> <p>NP Advantage (competitive) USE Competitive advantage</p> <p>NP Advantage (unfair competitive) USE Competitive advantage</p>

Figure 1. Thesaurus Excerpt with Legend

5.2. Initial DTD and Sample Instance

The initial DTD developed by one of the authors is [available](http://kensall.com/gov/glossary/glossary.dtd.txt) [http://kensall.com/gov/glossary/glossary.dtd.txt]. [Figure 2](#) shows an XML instance conforming to the DTD, illustrating a single term, "ontology". The instance depicts two `DefinitionSections` (author's terminology), each of which is a completely different definition of ontology (i.e., in a different scope). Since our knowledge of thesaurus standards was very limited at that point, the names of the elements were not particularly well-chosen. For example, based on the terminology presented in the previous section, `Term` should be `Concept`, `Name` should be `PreferredTerm`, `Usage` (and possibly `Concept` as well) should be `ScopeNote`, and so on.


```

<?xml version="1.0" encoding="UTF-8" ?>
<!-- ?xml-stylesheet href="glossary-sort-format3.xsl" type="text/xsl"? -->
<!DOCTYPE Glossary (View Source for full doctype...)>
- <Glossary>
- <Term id="ontology">
  <Name>ontology</Name>
  - <DefinitionSection>
    <Concept>semantic web</Concept>
    <Concept>knowledge management</Concept>
    <Definition>Defines the common words and concepts used to describe
      and represent an area of knowledge, and so standardizes the
      meanings. An ontology includes classes in the domains of interest,
      instances, relationships, properties and their values, functions of and
      processes involving the objects, and relevant constraints and
      rules.</Definition>
    <Source>Daconta, Obrst, Smith</Source>
    <Usage>An onontology can range from the simple notion of a taxonomy to
      a thesaurus, to a conceptual model, to a logical theory. [Daconta,
      Obrst, Smith]</Usage>
    <Synonym>classification system</Synonym>
    <RelatedTerm>taxonomy</RelatedTerm>
    <RelatedTerm>OWL</RelatedTerm>
  </DefinitionSection>
  - <DefinitionSection>
    <Concept>philosophy</Concept>
    <Definition>[sometimes "Ontology"] the metaphysical study of the
      nature of being and existence</Definition>
    <Source>WordNet</Source>
    <Usage>Both the ontology and manner of human existence are of
      concern to Existentialism.</Usage>
    <Synonym>metaphysics</Synonym>
  </DefinitionSection>
</Term>
</Glossary>

```

Figure 2. Instance of Initial DTD Showing One Term

5.3. XSLT-Generated Search Links

One aspect of this early approach that was particularly well received was an XSLT stylesheet that needed only a preferred term (i.e., a Name in our DTD) to generate links to various search engines. These links provide an easy way to check the current definition of a term against definitions found on the web, such as on WordNet or via Google define. [Visit Google and enter "define:ontology", for example.] Furthermore, the search links also function as a bootstrap process for term authors. Since nearly all of the elements in the DTD were optional, authors need only create an instance with the various Name elements they wish to define, apply the stylesheet, follow the generated links, and use the results to formulate their own definition.

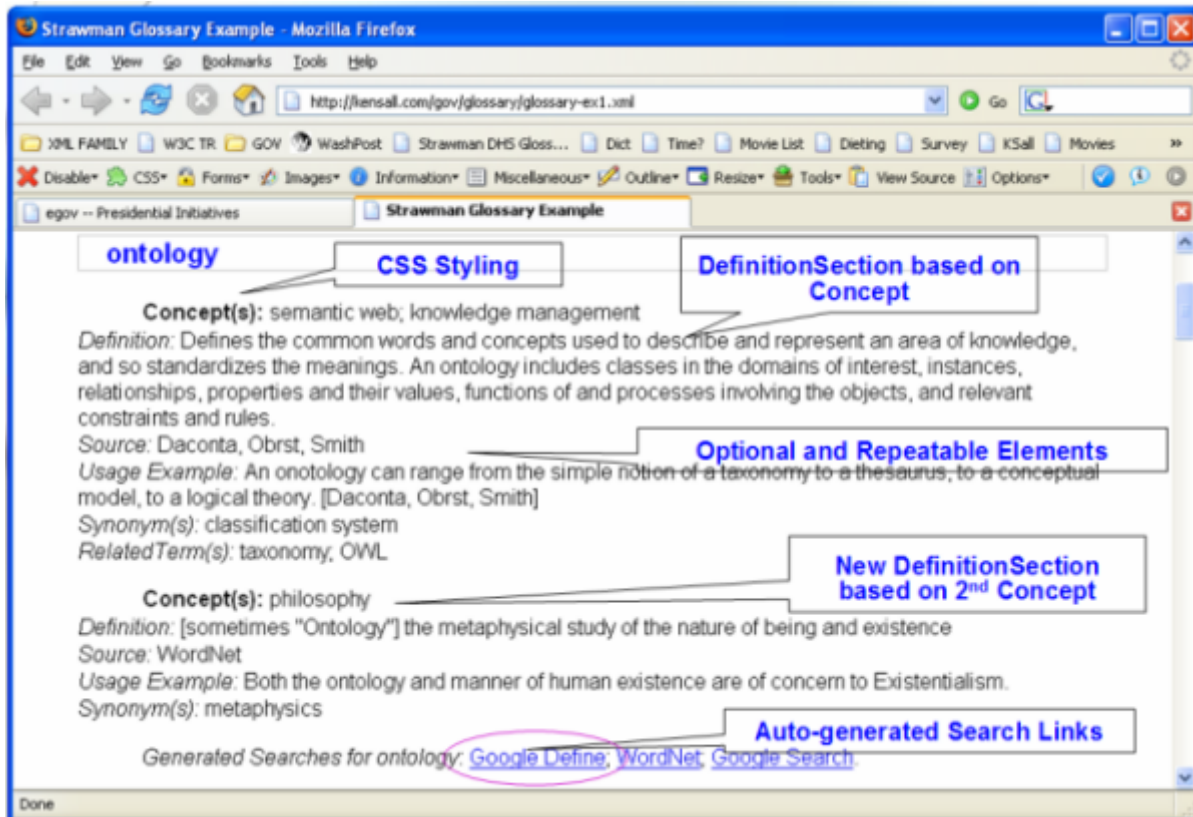


Figure 3. Term with Styling and Generated Search Links

Ultimately, we extended the stylesheet to generate up to fourteen links per concept, with links to AcronymFinder (if an acronym is identified), Wikipedia, Clusty, Clusty Gov [.gov and .mil], Google Uncle Sam [.gov and .mil], Google Define, Google, Merriam-Webster, W3C, W3Schools, Webopedia, WhatIs, WordNet, and ZVON. Results from Wikipedia and Google Define have been especially useful. See the [complete example](http://kensall.com/gov/glossary/glossary-ex1.xml) [http://kensall.com/gov/glossary/glossary-ex1.xml].

6. Thesauri Standards and Specifications

Based on feedback we received from one of the government working groups, we conducted two weeks of research of existing ISO and ANSI standards that relate to glossaries and thesauri. This section highlights the particular subset of the specifications we considered in the next phase of our work. As this is not our area of expertise, there are almost certainly other relevant standards that we have overlooked.

6.1. ISO 2788 - Developing a Thesaurus

ISO 2788:1986, Documentation - Guidelines for the establishment and development of monolingual thesauri [see [ISO-2788](#)], is a seminal specification from ISO/TC 46 (International Organization for Standardization Technical Committee). It provides very detailed guidance about how to select the indexing terms for a thesaurus, as well as how to express a variety of relationships. The standard explains all of the terminology discussed earlier (SN, USE, UF, BT, NT, RT, etc.) plus many others. Some of the main topics covered by ISO 2788 are:

- choice of singular vs. plural form

- choice of terms
- scope notes
- compound terms
- equivalent, hierarchical and associative relationships
- alphabetical, systematic and graphical displays

6.2. ISO 1087 - Vocabulary of Terminology

ISO 1087:2000, TERMINOLOGY WORK — VOCABULARY — Part 1: Theory and application [see [ISO-1087](#)], consists of vocabulary (normative) and concept diagrams (informative). The specification defines approximately 100 terms. Illustrative and paraphrased examples follow.

- Subject field (domain) - field of special knowledge
- Concept - unit of knowledge created by a unique combination of characteristics
- Characteristic - abstraction of a property of an object or of a set of objects
- Extension - set of objects to which a concept corresponds
- Intension - set of characteristics which make up the concept
- Hierarchical Relation
 - Generic Relation: vehicle and car (vehicle is a generalization of car)
 - Partitive Relation: week and day (day is part of a week)
- Associative Relation: baking and oven (experience tells us we bake in an oven)
- Extensional definition - enumerating all subordinate concepts under one criterion of subdivision (e.g., noble gases = {helium, neon, argon, krypton, xenon, radon})

Terminology work has 3 types of Designators (representation of a concept by a sign that denotes it)

- Symbol - remember the Formally-Known-As-Prince thing?
- Appellation - verbal designation of an individual concept
- Term - verbal designation of a general concept in a specific subject field; may have variants (i.e., alternate spellings)

ISO 1087 defines a number of kinds of terms, a sample of which include:

- Simple - one root (e.g., book)
- Complex - two or more roots (e.g., bookmaker, fault tolerance)
- Clipped term - abbreviation formed by truncating part of a simple term (e.g., flu for influenza, vet for veterinarian)
- Blend - formed by clipping and combining two separate terms (e.g., infomercial = information + commercial, ezine = electronic + magazine)

A few other concepts presented in ISO 1087 follow:

- Polysemy - one designation represents two or more concepts sharing certain characteristics (e.g., bridge: structure to carry traffic over a gap; dental plate; a card game)
- Homonymy - one designation represents two or more unrelated concepts (e.g., bark: sound made by dog; sailing vessel; outer part of a tree)
- Terminological dictionary - collection of terminological entries presenting information related to concepts or designations from one or more specific subject fields
- Vocabulary - terminological dictionary which contains designations and definitions from one or more specific subject fields
- Glossary - terminological dictionary which contains a list of designations from a subject field, together with equivalents in one or more languages [In English common language usage glossary can refer to a unilingual list of designations and definitions in a particular subject field.]

6.3. ISO 704 - Principles and Methods

ISO 704:2000, Terminology Work: Principles and methods [see [ISO-704](#)], is a product of Technical Committee ISO/TC 37, Terminology. This standard establishes basic principles and methods for preparing and compiling terminologies and describes the links between objects, concepts, and their representations through the use of terminologies. It borrows terms from ISO 1087-1:2000 (i.e., object, concept, characteristic, intension, extension, etc.).

ISO 704 explains the difference between essential vs. non-essential characteristics of a concept. Consider a pencil. Which of these characteristics is indispensable to understanding the concept: graphite is encased in wood, or one end may be sharpened to a point? The first characteristic is considered essential. However, a property may be an essential characteristic of a concept in one subject field but non-essential in another. There are also delimiting characteristics - essential characteristic that distinguishes one concept from another. One of the key principles stated in ISO 704:2000 is: "When modeling a concept system, one shall concentrate on the essential and delimiting characteristics." According to the specification, concept systems "model concept structures based on specialized knowledge of a field; [and] clarify the relations between concepts."

In addition to hierarchical relations (generic and partitive), the standard covers associative relations - thematic connection between concepts based on our experience with the objects. For example,

- Pencil case : pencil :: container : contained
- Writing : pencil :: activity : tool

6.4. ANSI/NISO Z39.19 - Construction, Format, and Management

[NISO](http://www.niso.org/standards/index.html) [http://www.niso.org/standards/index.html] (The National Information Standards Organization) describes *ANSI/NISO Z39.19-2003, Guidelines for the Construction, Format, and Management of Monolingual Thesauri*, as follows:

Z39.19 shows how to formulate descriptors, establish relationships among terms, and present the information in print and on a screen. Included are thesaurus maintenance procedures and recommended features for thesaurus management systems. Extensive examples, suggestions for further reading, and a detailed index complete this outstanding standard.

The main topics of Z39.19 are:

- scope, form, and choice of descriptors (single, compound, plurals, grammatical forms, abbreviations, acronyms, slang, jargon, capitalization, punctuation, etc.)
- compound terms as descriptors

- relationships (semantic linking, equivalence, hierarchical, associative)
- print display vs. screen display considerations
- thesaurus construction, maintenance, and management systems

This is the newest of the specifications we reviewed and it includes many of the same concepts and similar definitions as the ISO specs. In most respects, the topics covered by Z39.19 overlap and (in some cases) update material presented in ISO 2788. Note, however, that Z39.19 is an American standard rather than an international standard. On the other hand, Z39.19 is available at no charge, whereas most of the ISO standards mentioned in this paper have an associated fee. See [[Z39.19](#)].

6.5. Additional Specifications and Thesauri Sites

Other relevant specifications worth consideration when developing a thesaurus are:

- *ISO 12200:1999, Computer applications in terminology - Machine-readable terminology interchange format (MARTIF)* - Negotiated interchange; facilitates the interchange of termbase resources with conceptual data models of all different levels of complexity; presupposes a concept orientation rather than a word orientation [[MARTIF](#)].
- OLIF (Open Lexicon Interchange Format) -XML-compliant standard that can streamline the exchange of terminological and lexical data; used to interchange of data among lexbase resources from various machine translation systems; powerful but complex DTD [[OLIF](#)].
- SALT (Standards-based Access to multilingual Lexicons and Terminologies) - combines OLIF and NARTIF interchange formats [[SALT](#)].
- XLT (XML representation of Lexicons and Terminologies) [[XLT](#)].
- CLS Framework (Concept-oriented with Links and Shared references) [[CLS](#)].
- *ISO 15836:2003(E). Information and documentation - The Dublin Core metadata element set*; has a number of Relation refinements (conformsTo, hasFormat, hasPart, hasVersion, isFormatOf, isPartOf, isReferencedBy, isReplacedBy, isRequiredBy, isVersionOf, references, replaces, and requires) [[ISO-15836](#)].
- GlossXML - Proposed XML Format for Glossaries - a simplified, cross-platform approach based on a fairly simple DTD containing 19 elements [[GlossXML](#)].

Notable on-line thesauri worth investigating include:

- [WordNet](http://wordnet.princeton.edu/) [http://wordnet.princeton.edu/], a lexical database for the English language
- [GEMET thesaurus](http://www.eionet.eu.int/gemet/rdf) [http://www.eionet.eu.int/gemet/rdf] - EIONET (European Environment Information and Observation Network)
- [CSA-NBII](http://thesaurus.nbi.gov/SearchNBIIThesaurus/) [http://thesaurus.nbi.gov/SearchNBIIThesaurus/] (National Biological Information Infrastructure) Biocomplexity Thesaurus
- [MeSH](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=mesh) [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=mesh], NLM's controlled vocabulary used for indexing articles for MEDLINE/PubMed. MeSH terminology
- [CERES/NBII Thesaurus](http://ceres.ca.gov/thesaurus/) [http://ceres.ca.gov/thesaurus/]
- [World Bank Thesaurus](http://www2.multites.com/wb/) [http://www2.multites.com/wb/]

7. SKOS - W3C Working Draft

In February 2005, we first encountered SKOS (Simple Knowledge Organisation System), then an early Working Draft from the W3C. More specifically, SKOS is an emerging specification from the [Semantic Web Best Practices and Deployment Working Group](http://www.w3.org/2001/sw/BestPractices/) [http://www.w3.org/2001/sw/BestPractices/] of the W3C. At the time of this writing, the most recent SKOS documents were dated May 2005; since these are working drafts, it is likely that some details will change.

According to the SKOS Core Guide [[SKOS-Core](#)],

SKOS Core provides a model for expressing the basic structure and content of concept schemes (thesauri, classification schemes, subject heading lists, taxonomies, terminologies, glossaries and other types of controlled vocabulary). The SKOS Core Vocabulary is an application of the Resource Description Framework (RDF), that can be used to express a concept scheme as an RDF graph. Using RDF allows data to be linked to and/or merged with other RDF data by semantic web applications.

See [[SKOS](#)] for the SKOS home page. For a concise explanation of how to use SKOS for a thesaurus, as well as how it can be integrated with other RDF data such as Dublin Core, see [[SKOS-Quick](#)]. Readers wishing a less technical introduction to SKOS might enjoy [[SKOS-Intro](#)] available at XML.com.

SKOS Classes	SKOS Properties
CollectableProperty	<i>altLabel - implemented in pilot (see text)</i>
Collection	altSymbol
<i>Concept - implemented in pilot</i>	<i>broader - implemented in pilot</i>
ConceptScheme	changeNote
OrderedCollection	<i>definition - implemented in pilot</i>
	editorialNote
	<i>example - implemented in pilot</i>
	hasTopConcept
	hiddenLabel
	historyNote
	inScheme
	isPrimarySubjectOf
	isSubjectOf
	member
	memberList
	<i>narrower - implemented in pilot</i>
	<i>prefLabel - implemented in pilot</i>
	prefSymbol
	primarySubject
	privateNote
	publicNote
	<i>related - implemented in pilot</i>
	<i>scopeNote - implemented in pilot</i>
	semanticRelation
	<i>subject - implemented in pilot</i>
	subjectIndicator

Figure 4. SKOS Vocabulary

Figure 4 shows the current SKOS Vocabulary, consisting of 5 Classes and 26 Properties, with an indication of which aspects have been implemented in our pilot to date. With regard to the `altLabel` element, however, we discourage using this to indicate acronyms and abbreviations, in contrast to the W3C working draft, because we want to be able to process acronyms and abbreviations separately from variant spellings and labels (e.g., to generate acronym lists). Note the SKOS convention of UpperCamelCase for Classes and lowerCamelCase for Properties. See [SKOS-Vocab] and [SKOS-Core] for details. Examples appear in later sections of this paper.

Using a visualization of an RDF Graph (Subject-Predicate-Object) in which Subject and Object are nodes represented by circles and the Predicate is the labeled arc, we can illustrate several of the key SKOS Properties in Figure 5, patterned after a similar illustration in [SKOS-Quick]. Each circle represents a `skos:Concept` and each arc is a Property identified by its label. We have color coded the concepts in our diagram as follows:

- black is the central concept with the `skos:prefLabel` "bird" (i.e., the black circle denotes the concept we call "bird")

- dark blue circles are concepts that are `skos:narrower` than the "bird" concept
- light blue circles are concepts that are `skos:broader` than the "bird" concept
- white circles represent concepts that are `skos:related` to the "bird" concept

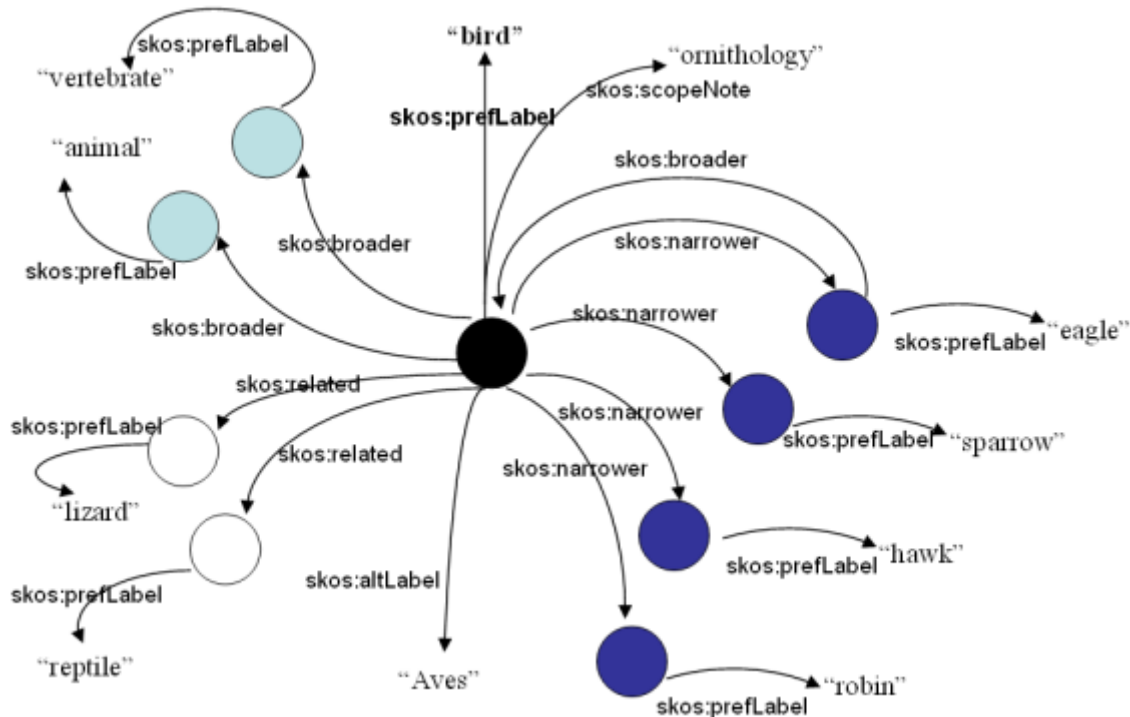


Figure 5. RDF Graph Visualization of Bird Relationships in SKOS

From [Figure 5](#), some of the many statements we can make include:

- An alternate label (`skos:altLabel`) for "bird" is "Aves".
- The concepts with the preferred label "vertebrate" and "animal" are broader than the concept with the preferred label "bird".
- There are four specializations of birds listed ("robin", "hawk", "sparrow" and "eagle"), each indicated as `skos:narrower` than "bird".
- The concepts "lizard" and "reptile" are `skos:related` to the "bird" concept in some way.
- Among various concepts which might have the `skos:prefLabel` of "bird", the one illustrated is constrained to ornithology, according to `skos:scopeNote`. This distinguishes the concept from "bird", such as in the informal term for a (young) woman.

The properties `skos:broader` and `skos:narrower` are inverses of each other. For example, the concept with the `skos:prefLabel` "bird" and the concept with `skos:prefLabel` "eagle" have two arcs shown, indicating that a narrower term for "bird" is "eagle" and a broader term for "eagle" is "bird". (Due to the inverse relationship, one of the arcs is technically redundant.) Both `skos:broader` and `skos:narrower` are transitive properties as well. For example, if "eagle" is narrower than "bird", and "animal" is broader than "bird" (i.e., "bird" is narrower than "an-

imal"), then it follows that "eagle" is narrower than "animal" (and "animal" is broader than "eagle"). Note that it is precisely these kinds of conclusions about semantic relationships that makes a thesaurus preferable to a glossary.

8. Our SKOS Element Subset and Extensions

For the purposes of our pilot, we decided to implement a subset of SKOS elements and to introduce several elements of our own. In [Table 1](#), elements shown in uppercase (COI, ABBREVIATION_OR_ACRONYM, and SOURCE) are not part of the SKOS Core Vocabulary Specification Working Draft. In March 2005, one of the authors created a [simple XML Schema](http://kensall.com/gov/glossary/thesaurus.xsd.txt) [http://kensall.com/gov/glossary/thesaurus.xsd.txt] that implements all of the elements shown in the figure (except for COI) with suggestions from Judith Newton (<jnewton@ashcomp.com>).³

³Actually, there are several additional elements in the schema such as `skos:changeNote`, `skos:editorialNote`, `skos:semanticRelation`, and our own `REVIEW` element which have not been incorporated into our prototype at this stage.

Element	Definition	Origin
Concept	Represents an abstract idea or notion; a unit of thought. This is essentially a term with possibly related terms.	SKOS Core
COI	Each Community of Interest represents a collection of concepts. The idea is to permit separate concept management for each COI, although ultimately linkages across COIs would be encouraged.	Sall
prefLabel	Preferred name for this concept. This is typically thought of as the name of the term, sometimes called the designator in ISO specifications. NOTE: In SKOS, no two concepts in the same concept scheme may have the same value.	SKOS Core
altLabel	Alternate name for this concept. Spelling variants and irregular plural/singular forms may be included among the alternative labels for a concept. NOTE: <i>In a departure from the SKOS Core Vocabulary Specification Working Draft, term authors should not place abbreviations or acronyms here.</i>	SKOS Core
ABBREVIATION or ACRONYM	Although the SKOS Core Vocabulary Specification Working Draft suggests that altLabel may include abbreviations and acronyms for the concept, our design calls for placing abbreviations and acronyms in this separate element. We believe this will enable easy filtering or querying to construct abbreviation and acronym lists that often appear in government documents.	Sall
definition	A statement or formal explanation of the meaning of a concept. Although this is not technically required, most instances are expected to include a definition.	SKOS Core
subject	A concept that is the subject of the current term. This may be one of the "broader" terms or it could be related to the scopeNote. Ex: Subject of Toyota Camry is automobiles.	SKOS Core
scopeNote	A note that helps to clarify the meaning of a concept. This could be contextual information, a particular domain, or some other constraint to bound the scope of the concept, perhaps to distinguish it from the same term (prefLabel) used in a different scope. While ISO 2788 allows definitions to appear in the scopeNote, we are following the SKOS approach of using the separate definition element for that purpose.	SKOS Core
SOURCE	Source of the definition. Official document names and URLs are preferred, but specific names of people or agencies are acceptable. There is no enforced pattern for this data presently, but we expect a future version will specify the bibliographic citation conventions.	Sall
example	An example of the use of the concept, such as in a sentence.	SKOS Core
narrower	A concept that is more specific in meaning. Narrower concepts are typically rendered as children in a concept hierarchy (tree). Ex: Bald eagle is narrower than eagle; eagle is narrower than bird.	SKOS Core
broader	A concept that is more general in meaning. Broader concepts are typically rendered as parents in a concept hierarchy (tree). Ex: Bird is broader than eagle; eagle is broader than bald eagle.	SKOS Core
related	A concept with which there is an associative semantic relationship. Ex: Nest, flying, and wings are all concepts that are related to the concept bird	SKOS Core

Table 1. Implemented SKOS Subset and Extensions

9. Pilot Solution

At the time of this writing, our two-man team (the authors) has worked part time on a proof-of-concept for approximately two months. The pilot, dubbed **SKOSSaurus** [<http://skosaurus.rrecktek.com/>] by one author, is an open source solution based on perl, Java, SOAP, Kowari.

9.1. Pilot Environment

The environment used to demonstrate SKOSSaurus is multilayer and somewhat complicated. First, at the foundation level, there is a host operating system that runs machine emulator software called VMware. The machine emulator software simulates a machine that itself has an operating system. In turn, this operating system has a host of supporting software that provides the SKOSSaurus application. See [[VMware](#)].

The host operating system for SKOSSaurus is Microsoft Windows XP with Service Pack 2 running on a Dell Latitude D800 (1.69GHz) with 1G of RAM. The Windows XP host runs VMware 5.0 build 13124. VMware is used to emulate a machine onto which the Solaris X86 operating system version 10 is installed. This is referred to as the guest operating system which runs the SKOSSaurus system, consisting of Perl version 5.8.7 and various Perl modules, Java version 1.4.2.08, and the Kowari server 1.1.0 Pre2. [[Kowari](#)]

9.2. Use Cases

For our proof-of-concept pilot, we chose to implement only the most essential of the possible use cases:

- Term Entry via Web Form
- File Upload of Excel Spreadsheet
- File Upload of XML or SKOS
- Query of Term Data Store

These use cases were selected because we felt that the ability to ingest Excel spreadsheets was of paramount importance for our potential user community. We felt it was crucial to have an easy way for term authors who were largely unfamiliar with XML (not to mention RDF and SKOS) to define terms and express relationships in a natural manner. Therefore, we provided a spreadsheet with the 11 columns pictured in [Figure 6](#) and [Figure 7](#) and described in [Table 1](#).

9.3. Assumptions

The assumptions presented in this section were considered acceptable limitations for the purpose of a pilot. Most if not all of these issues would be addressed in a production system.

- No file locking - It is possible although highly unlikely that users coming from the same IP address at the exact same second, could experience a loss of data.
- Bad characters - The system does not check for bad characters in submitted or uploaded data. In particular, no attempt is made to parse for character entities or for characters outside the default XML encoding. For example, Microsoft Office special quotes and long dashes are not handled properly and must be avoided.
- UNIX only - The system is designed to run under Unix (Solaris) or a Unix variant (Linux).
- Browser limitations - The upload feature only works with HTTP 1.1 compliant browsers. This should not be a problem as all reasonably recent browsers meet this requirement.

-
- Limited tagset - Only SKOS Core and Sall elements are supported.
 - Limited delimiters - The only permissible CVS delimiters are commas, tabs, or pipe.
 - Disk Space - Neither disk space checking nor file upload size limits are enforced.
 - Server requires a graceful shutdown - An unanticipated server shutdown can cause a variety of problems. Each of the steps in the pilot are expected to complete successfully. If there is an interruption, there is no graceful recovery mechanism.
 - No field validation - JavaScript field checking is not implemented for the web forms.
 - No Backups - Regular backups of the system are not yet supported.
 - Whitespace preservation - There is no preservation of white space. Multiple spaces are likely to become a single space.
 - Single Entry - Web form supports only one record (concept) at a time. However, CSV upload permits ingesting a virtually unlimited number of concepts in one upload.

9.4. Spreadsheet Format and Birds Example

SKOSaurus supports the uploading of a Microsoft Excel spreadsheet that has been exported in CSV (Comma Separated Values) format. This method of data entry permits the bulk loading of data and is the preferred means of providing a data source to the pilot system.

Figure 6 and Figure 7 illustrate the left and right halves of a sample spreadsheet. Each row corresponds to a single concept. Of the eleven columns (`prefLabel`, `altLabel`, `ABBREVIATION_OR_ACRONYM`, `definition`, `subject`, `scopeNote`, `SOURCE`, `example`, `narrower`, `broader`, and `related`), only `prefLabel` is actually required. However, the more cells a term author fills in, the more semantics can be applied to the resultant concept once it is converted to SKOS.

Several important conventions must be followed by term authors when completing the spreadsheet:

- The heading row should not be removed or modified. (We recommend a spreadsheet with the heading row locked.)
- Since several elements are repeatable, we expect authors to use semi-colon to indicate iteration. More about this later.
- Although conversion to CSV handles commas within a cell by surrounding the entire cell contents in double quotes, a limitation in our pilot parser requires the author to use the pipe symbol ("|") instead of a comma within a cell. (Of course, this limitation would not be acceptable in a production system.)
- Any number of rows can be included, but there must be no blank rows or separator rows.
- When the author has completed the spreadsheet, he selects `File > Save As` from the Microsoft Excel menubar and saves the files as Comma Separated Values (*.csv).

9.5. Data Submission and Conversion to SKOS

The top level menu of SKOSaurus consists of the choices:

- Manage COIs [Communities of Interest]
- Upload a CSV file

prefLabel	altLabel	ABBRE	definition	subject	scopeNote
bird	Aves		warm-blooded egg-laying vertebrates characterized by feathers and forelimbs modified as wings	ornithology	ornithology
bird	chick		informal term for a (young) woman	slang	slang; more likely to be used in the UK
shuttlecock	birdie; cock; shuttle; bird		A shuttlecock is a high-drag projectile used in the sport of badminton. It has an open conical shape with a rounded head at the apex of the cone traditionally made of cork and a skirt traditionally of sixteen overlapping goose feathers.	sports	This term is applies only to the sport of badminton.
birdie	birdy		A score on an individual hole that is one stroke below par.	golf	This term applies only to golf.
eagle			any of various large keen-sighted diurnal birds of prey noted for their broad wings and strong soaring flight	ornithology	This terms refers to ornithology.
eagle			a score of two strokes under par on a hole	golf	This term applies only to golf.
Eagles			The Eagles are an American rock music group that originally came together in Los Angeles California in the early 1970s.	music	This applies only to the rock music group.
Eagles	Philadelphia Eagles		Philadelphia football team	sports	This is a football team.
robin	American Robin		The American Robin (Turdus migratorius) is a	ornithology	This terms refers to

Figure 6. Birds Excel Spreadsheet, Left Half

- Enter values [via a Web Form]
- Kowari based Search

prefLabel	SOURCE	example	narrower	broader	related
bird	WordNet [http://wordnet.princeton.edu/] and http://en.wikipedia.org/wiki/Bird	A bird is an animal with feathers. All birds have wings but not all birds can fly. Ostriches emus and penguins are birds that can't fly. Parrots pelicans and wedge-tailed eagles are birds that can fly. All birds hatch from eggs. All birds are warm blooded.	eagle; sparrow; hawk; robin; kiwi	vertebrate; chordate; animal	birding; birdwatching; reptile; lizard
bird	http://wordnet.princeton.edu/	The young man whistled as the blonde bird walked by.		female	dame; doll; wench; skirt
shuttlecock	http://en.wikipedia.org/wiki/Birdie			badminton	
birdie	http://golf.about.com/cs/golfterms/g/bldef_birdie.htm				par; bogey; even; eagle
eagle	WordNet [http://wordnet.princeton.edu/] and http://en.wikipedia.org/wiki/Eagle		bald eagle	bird	raptor
eagle	WordNet [http://wordnet.princeton.edu/]				
Eagles	http://en.wikipedia.org/wiki/Eagles	The Eagles compilation Their Greatest Hits (1971-1975) with US sales of 28 000 000 is the best selling album of all time.	Glenn Frey; Don Henley; Joe Walsh; Timothy Schmit	California Rock	Linda Ronstadt; Poco; James Gang
Eagles	http://www.philadelphiaeagles.com/default.jsp	The Philadelphia Eagles are part of the NFL.	Donovan McNabb	NFL	Washington Redskins
robin	http://en.wikipedia.org/wiki/American_Robin	The American Robin has gray upperparts and head and orange underparts usually		thrush; bird	blackbird; European Robin

Figure 7. Birds Excel Spreadsheet, Right Half

The "Manage COIs" choice lets the user create or delete four separate slices of the data store. This is purely a management function and will be handled differently in the production system. The other three menu choices are discussed at length in the following sections.

9.5.1. Data Entry via a Web Form

One mechanism for entering data into the SKOSaurus system is by means of a web form. The SKOSaurus web form consists of a series of field names followed by text entry fields, as shown in [Figure 8](#). As explained earlier, all of the fields correspond to elements from the SKOS Core Vocabulary except for COI, abbreviation_or_acronym, and source. A limitation of the current system is that fields that can be repeated in the spreadsheet are not repeatable in the data entry form.

The screenshot shows a web browser window titled "SKOS DEMO - Mozilla Firefox" with the address bar showing "http://skosaurus.rrecktek.com/cgi-bin/hoard.pl". The form contains the following fields:

- concept:
- COI: IC MWG IC EA CAF DRM
- prefLabel:
- altLabel:
- abbreviation or acronym:
- definition:
- subject:
- scopenote:
- source:
- example:
- narrower:
- broader:
- related:

A "Submit" button is located at the bottom left of the form.

Figure 8. SKOSaurus Web Form for Entering a Single Concept

For the sake of consistency, each field name matches the corresponding SKOS element as well as the HTML name, and the underlying variable in the Perl program that creates and stores the file after submission. For example, the underlying HTML corresponding to the `prefLabel` field in [Figure 8](#) is:

```
<input type="text" name="prefLabel" size="60" >
```


The advantage of this approach is that the form need not reside on the system that hosts SKOSaurus. Any web server could conceivably host this form since it is really just a web page. Each COI could have its own customized version of the web form. The most important aspect is the underlying POST line:

```
<form method="post" action="http://skosaurus.rrecktek.com/cgi-bin/hoard.pl"
enctype="application/x-www-form-urlencoded">
```

Any web form that is posted to hoard.pl will result in correctly formatted SKOS, provided it contains name value pairs for the eleven elements shown in the figure. While some field validation could be accomplished using JavaScript imbedded in the form, additional validation would still need to occur at the server level since there is nothing to prohibit POSTs with non-validated data.

After the form is submitted, a Perl CGI program called hoard.pl is invoked. The Perl script takes the list of name/value pairs transmitted via the web form and creates the SKOS RDF statements that correspond to the values that were submitted. This is the place where field validation should occur a second time to prohibit unacceptable characters (such as the wrong kind of quotes or dashes). The lines of Perl that capture values from the submitted form set a variable with lines that look like:

```
$prefLabel = $query->param('prefLabel');
```

We have chosen to use the same variable names in the Perl script as their corresponding SKOS property names. The resultant SKOS statement for this single text entry field is:

```
<skos:prefLabel xml:lang="en">bird</skos:prefLabel>
```

9.5.2. Uploading CSV Data

However, the more efficient way to ingest concepts into SKOSaurus is by creating a spreadsheet, saving it as CSV, and uploading it to our server. For the purposes of demonstration, we are using a simple web based file upload capability. Term authors visit the upload page and press the [Browse] button. A window opens and allows them to navigate their local file system to find the file they want to upload. Once the file is selected, the user presses the [Open] button to set the pathname to the file. The user is free to change the field delimiter from the default value of comma to either a pipe or tab via a radio button. It is also necessary to specify with which of the COIs the stored SKOS is to be associated.⁴

After the desired filename, field delimiter, and COI are indicated, the user presses the [Upload] button to transmit the file to the web server's incoming queue. The server then provides the user feedback with a message containing links to the uploaded CSV and the converted SKOS.

Upload received. See the upload or the resulting SKOS.

A second message indicates that the Kowari MetaStore server successfully loaded the resulting SKOS:

```
Successfully loaded 55 statements from file:/var/apache2/htdocs/incoming/csv-
1126091939.txt into rmi://skos-demo/server1#icmwg
```

This particular message indicates the number of SKOS statements that were generated and into which COI model the data was loaded (e.g., "icmwg"). Note that for each row in the spreadsheet, a number of SKOS statements is generated. At least one statement is generated for each cell, and cells containing the iteration character (discussed in the next section) result in multiple SKOS statements. [Figure 9](#) shows the SKOS statements generated for the first two rows of the spreadsheet shown in [Figure 6](#) and [Figure 7](#).

⁴Choice of COI is completely immaterial in the pilot, but of real value in the production system. The four COIs depicted in the interface are IC MWG (Intelligence Community Metadata Working Group), ICEA (Intelligence Community Enterprise Architecture), CAF (Chief Architects Forum), and DRM (Data Reference Model). It should be noted that none of these groups has committed to using our SKOSaurus system as of this writing.

```

- <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:skos="http://www.w3.org/2004/02/skos/core#">
- <skos:Concept rdf:about="http://skos.rrecktek.com/icmwg/1126816675-470184#concept">
  <skos:prefLabel xml:lang="en">bird</skos:prefLabel>
  <skos:altLabel xml:lang="en">Aves</skos:altLabel>
  <skos:definition xml:lang="en">warm-blooded egg-laying vertebrates characterized by feathers and
    forelimbs modified as wings</skos:definition>
  <skos:scopenote xml:lang="en">ornithology</skos:scopenote>
  <skos:source xml:lang="en">WordNet [http://wordnet.princeton.edu/] and
    http://en.wikipedia.org/wiki/Bird</skos:source>
  <skos:subject
    rdf:resource="http://skosaurus.rrecktek.com/bootstrap/icmwg/ornithology#concept" />
  <skos:example xml:lang="en">A bird is an animal with feathers. All birds have wings but not all birds can
    fly. Ostriches, emus and penguins are birds that can't fly. Parrots, pelicans and wedge-tailed
    eagles are birds that can fly. All birds hatch from eggs. All birds are warm blooded.</skos:example>
  <skos:broader
    rdf:resource="http://skosaurus.rrecktek.com/bootstrap/icmwg/vertebrate#concept" />
  <skos:broader rdf:resource="http://skosaurus.rrecktek.com/bootstrap/icmwg/chordate#concept" />
  <skos:broader rdf:resource="http://skosaurus.rrecktek.com/bootstrap/icmwg/animal#concept" />
  <skos:narrower rdf:resource="http://skosaurus.rrecktek.com/bootstrap/icmwg/eagle#concept" />
  <skos:narrower rdf:resource="http://skosaurus.rrecktek.com/bootstrap/icmwg/sparrow#concept" />
  <skos:narrower rdf:resource="http://skosaurus.rrecktek.com/bootstrap/icmwg/hawk#concept" />
  <skos:narrower rdf:resource="http://skosaurus.rrecktek.com/bootstrap/icmwg/robin#concept" />
  <skos:narrower rdf:resource="http://skosaurus.rrecktek.com/bootstrap/icmwg/kiwi#concept" />
  <skos:related rdf:resource="http://skosaurus.rrecktek.com/bootstrap/icmwg/birding#concept" />
  <skos:related
    rdf:resource="http://skosaurus.rrecktek.com/bootstrap/icmwg/birdwatching#concept" />
  <skos:related rdf:resource="http://skosaurus.rrecktek.com/bootstrap/icmwg/reptile#concept" />
  <skos:related rdf:resource="http://skosaurus.rrecktek.com/bootstrap/icmwg/lizard#concept" />
</skos:Concept>
- <skos:Concept rdf:about="http://skos.rrecktek.com/icmwg/1126816675-740196#concept">
  <skos:prefLabel xml:lang="en">bird</skos:prefLabel>
  <skos:altLabel xml:lang="en">chick</skos:altLabel>
  <skos:definition xml:lang="en">informal term for a (young) woman</skos:definition>
  <skos:scopenote xml:lang="en">slang; more likely to be used in the UK</skos:scopenote>
  <skos:scopenote xml:lang="en">slang; more likely to be used in the UK</skos:scopenote>
  <skos:source xml:lang="en">http://wordnet.princeton.edu/</skos:source>
  <skos:subject rdf:resource="http://skosaurus.rrecktek.com/bootstrap/icmwg/slang#concept" />
  <skos:example xml:lang="en">The young man whistled as the blonde bird walked by.</skos:example>
  <skos:broader rdf:resource="http://skosaurus.rrecktek.com/bootstrap/icmwg/female#concept" />
  <skos:related rdf:resource="http://skosaurus.rrecktek.com/bootstrap/icmwg/dame#concept" />
  <skos:related rdf:resource="http://skosaurus.rrecktek.com/bootstrap/icmwg/doll#concept" />
  <skos:related rdf:resource="http://skosaurus.rrecktek.com/bootstrap/icmwg/wench#concept" />
  <skos:related rdf:resource="http://skosaurus.rrecktek.com/bootstrap/icmwg/skirt#concept" />
</skos:Concept>
- <skos:Concept rdf:about="http://skos.rrecktek.com/icmwg/1126816675-518179#concept">
  <skos:prefLabel xml:lang="en">shuttlecock</skos:prefLabel>
  <skos:altLabel xml:lang="en">birdie</skos:altLabel>

```

Figure 9. First Two Rows of Birds Spreadsheet Converted to SKOS

Other upload options we are exploring (in addition to the CSV file) include properly formed SKOS documents or XML documents adhering to a particular Microsoft XML Schema (e.g., `xmlns:od="urn:schemas-microsoft-com:officedata"`). In these cases, the field delimiter would be ignored since it is irrelevant.

9.5.3. Iteration Character

The CSV data upload facilitates concept loading *en masse*. To accomplish this, the system supports the notion of an iteration character which is presently a semicolon. If a semicolon appears in a field of CSV upload, the system generates another SKOS statement for each semicolon delimited value encountered. Consider the first row of the birds spreadsheet. The "narrower" cell contains a string with five semicolon-separated terms:

eagle; sparrow; hawk; robin; kiwi

This signals the generator to create five separate `skos:narrower` statements as shown in Figure 10 (which are a subset of the lines from Figure 9):

```
<skos:narrower rdf:resource="http://skosaurus.rrecktek.com/bootstrap/icmwg/eagle#concept" />
<skos:narrower rdf:resource="http://skosaurus.rrecktek.com/bootstrap/icmwg/sparrow#concept" />
<skos:narrower rdf:resource="http://skosaurus.rrecktek.com/bootstrap/icmwg/hawk#concept" />
<skos:narrower rdf:resource="http://skosaurus.rrecktek.com/bootstrap/icmwg/robin#concept" />
<skos:narrower rdf:resource="http://skosaurus.rrecktek.com/bootstrap/icmwg/kiwi#concept" />
```

Figure 10. Five Statements Generated by Iteration of the Narrower Cell

This affords the term author an easy and fairly natural way to express repetition within the constraints of a cell in the original spreadsheet.

9.6. Bootstrapping URIs

Since a strength of the SKOSaurus system lies in the ability to relate one term to another, this can pose a challenge at the inception of the system because there are no existing terms to which to relate. Therefore, the SKOSaurus system employs a bootstrapping strategy that allows a term author to relate a term to broader and narrower terms that may not yet exist in the data store, or even if they do exist, the author may be unaware of their existence.

Bootstrapping allows authors to use string literals in the `subject`, `narrower`, `broader`, and `related` fields instead of the URIs that SKOS actually requires. During the process of converting data to SKOS, the system determines whether the `subject`, `narrower`, `broader`, and `related` values are indeed legal URIs. If the values are string literals, the system generates a query behind the scenes for entries that have the corresponding `prefLabels` corresponding to the string literal. If the `prefLabel` exists, then the system substitutes the corresponding URI in place of the string literal in the generated SKOS statement.

For example, this input contains literals that should become URIs:

```
<skos:broader rdf:resource=
  "http://skosaurus.rrecktek.com/bootstrap/icmwg/animal#concept" />
<skos:narrower rdf:resource=
  "http://skosaurus.rrecktek.com/bootstrap/icmwg/eagle#concept" />
```

is changed to:

```
<skos:broader rdf:resource=
  "http://skosaurus.rrecktek.com/icmwg/1126816680-234516#concept" />
<skos:narrower rdf:resource=
  "http://skosaurus.rrecktek.com/icmwg/1126811175-67453#concept" />
```

where the numeric URIs are the result of the lookup and indeed identify the concepts with `prefLabels` "animal" and "eagle", respectively.

If input is via a web form, the user is given feedback that this substitution has occurred. If there is no corresponding entry with the right `prefLabel` in the knowledge base, SKOSaurus creates an impoverished entry for the broader or narrower term and marks it with a `SOURCE` of "bootstrap". (In the prototype SKOSaurus system, bootstrapping only occurs for CSV upload or web form data entry.)

9.7. Committing Data to Persistent Storage

Once the conversion to SKOS is completed, the SKOS file is stored in the database that serves as a backend for the system. The effort needed to move RDF into the data store is lessened because the server can work with native RDF. Our pilot SKOSaurus application uses an open source data store called Kowari Metastore by Tucana Technologies [Kowari]. According to the project homepage, "The Kowari Metastore is an Open Source, massively scalable, transaction-safe, purpose-built database for the storage and retrieval of metadata." Working with it is both exciting and straight forward. Kowari Metastore functionality is accessible through SOAP. Simply put, SOAP codifies the use of XML, generally over HTTP, to allow a client to invoke methods on a server. In this pilot, the SKOSaurus client and Kowari server happen to reside on the same machine. In general, however, this is not the case; when the client and server are on different machines, the power and scalability of the approach are apparent.

The data storage process begins when a SOAP client opens a connection to the Kowari Metastore and loads the SKOS data into a model corresponding to the selected COI. The steps are minimal as the following code fragment demonstrates:

```
#perl needs the use the right library
use SOAP::Lite;

## define four variables
# tell the program where to find the server
$serverEnd =
"http://skos-demo:8080/webservices/services/ItqlBeanService";
#use the correct command syntax to enter data
$iTQLcommand =
"load <file:$filename> into <rmi://skos-demo/server1#$coi>";
# tell the client program which file has the data
$filename=( $opt_f );
#which Community of Interest (model) are we working with
$COI=( $opt_c );

#now make the call to the subroutine.
print soapProcess($serverEnd, $iTQLcommand);

# the simple subroutine
sub soapProcess {
($serverSoapEndpoint, $iTQLcommand) = @_;
$resultString = SOAP::Lite
-> uri($serverSoapEndpoint)
-> proxy($serverSoapEndpoint)
-> executeQueryToString($iTQLcommand)
-> result;
return $resultString;
}
```

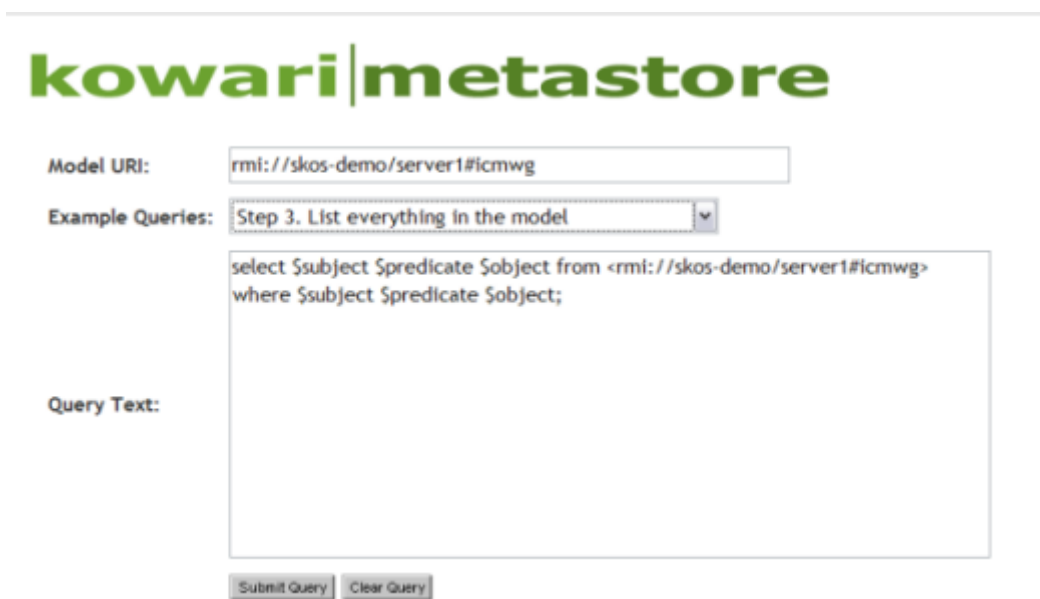
With less than 20 lines of Perl code, the file is added to the Kowari Metastore using a SOAP request. The elegance of this approach is that all interactions with the Kowari Metastore server are done in a machine independent and platform independent manner. This very versatile approach would support data entry from SOAP clients on multiple machines in mixed environments.

Weak security protection and access restrictions could easily be ameliorated by adding basic authentication to the SOAP call. The use of basic authentication over HTTP suffers because the username and password travel as clear text and are only protected through a Base64 encoding. Further, because of the stateless nature of the HTTP protocol, the

requestor must authenticate each time there is a need to access a protected resource. A slightly better approach to security is to use digest authentication. This does not preclude the ability to leverage the stronger protection of SSL/TLS if other HTTPS aware infrastructure were used. The point here is that there are several options for adding security to the transport layer used in the pilot, and each will integrate readily.

9.8. Querying the Data Store

In the current pilot, all queries to the data store are done through the default querying mechanism provide by Kowari, which again uses SOAP. The query screen has a text field at the top in which to specify the model that is to be queried. The user replaces the string `#sampledata` with the name of their selected COI. Next, the user can choose from a list of sample queries such as "List everything in the model", or construct a customized query. Once a query choice is selected, the large text area is populated with the actual query that will be executed when the [Submit Query] button is pressed. See [Figure 11](#) in which we have set the model to the COI `#icmwg`.



kowari|metastore

Model URI:

Example Queries:

Query Text:

```
select $subject $predicate $object from <rmi://skos-demo/server1#icmwg>
where $subject $predicate $object;
```

Figure 11. Kowari Form with Query Selected (Before Submitting)

[Figure 12](#) shows the first few lines of the query result in the default Kowari interface, in which activating any green link causes another query to be sent to the data store.

The screenshot shows the 'ari|metastore' interface. At the top, there is a text input field containing the URI 'rmi://skos-demo/server1#icmwg'. Below it is a dropdown menu with the text 'Step 3. List everything in the model'. A large text area contains the SPARQL query: 'select \$subject \$predicate \$object from <rmi://skos-demo/server1#icmwg> where \$subject \$predicate \$object;'. Below the query area are two buttons: 'Submit Query' and 'Clear Query'. Below the buttons, it says '(1 query, 0.989 seconds)'. The query result is displayed as a table with three columns: 'predicate' and 'object'. The first row shows a concept URI, a predicate URI, and a concept URI. The second row shows the same concept URI, a predicate URI, and the string 'bird'. The third row shows the same concept URI, a predicate URI, and the string 'Aves'. The fourth row shows the same concept URI, a predicate URI, and the string 'warm-blooded egg-laying vertebrates characterized'.

	predicate	object
tek.com/icmwg/1126622111-380957#concept	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2004/02/skos/core#Concept
tek.com/icmwg/1126622111-380957#concept	http://www.w3.org/2004/02/skos/core#prefLabel	"bird"
tek.com/icmwg/1126622111-380957#concept	http://www.w3.org/2004/02/skos/core#altLabel	"Aves"
tek.com/icmwg/1126622111-380957#concept	http://www.w3.org/2004/02/skos/core#definition	"warm-blooded egg-laying vertebrates characterized"

Figure 12. Kowari Query Result: Bird Fragment

If we compare the query result to the SKOS generated from the spreadsheet shown in [Figure 9](#), the correspondence to the first few lines is apparent.

```
<skos:Concept rdf:about=
  "http://skos.rrecktek.com/icmwg/1126622111-380957#concept">
  <skos:prefLabel xml:lang="en">bird</skos:prefLabel>
  <skos:altLabel xml:lang="en">Aves</skos:altLabel>
  <skos:definition xml:lang="en">
  warm-blooded egg-laying vertebrates characterized by feathers
  and forelimbs modified as wings</skos:definition>
  <-- etc. -->
```

If we were to then click on the green word "bird" appearing in the object column of [Figure 12](#), it would trigger this query:

```
select $subject $predicate $object from <rmi://skos-demo/server1#icmwg>
  where $subject $predicate 'bird'
```

In other words, find all subjects that have "bird" as the object, regardless of the predicate. The result of this query is shown in [Figure 13](#). It tells us there are three different concepts in the data store for which "bird" is the value of some SKOS property. Refer back to the spreadsheet in [Figure 6](#) to confirm that "bird" is the `prefLabel` in two cases (the first two rows) and is also one of the `altLabel` values in the third row.

ari|metastore

rmi://skos-demo/server1#icmwg

Select a query..

Submit Query Clear Query

{1 query, 0.063 seconds)

select \$\$Subject \$Predicate from <rmi://skos-demo/server1#icmwg> where \$\$Subject \$Predicate 'bird';

	Predicate
ek.com/icmwg/1126622111-380957#concept	http://www.w3.org/2004/02/skos/core#prefLabel
ek.com/icmwg/1126622111-815694#concept	http://www.w3.org/2004/02/skos/core#prefLabel
ek.com/icmwg/1126622111-944102#concept	http://www.w3.org/2004/02/skos/core#altLabel

Figure 13. Kowari Query Result: All Bird Objects

If we then click on the third concept shown in [Figure 13](#) which is on the row with the `altLabel` predicate, we are effectively asking to see all of the predicates and objects of this concept. The query result appears in [Figure 14](#). The retrieved concept is the one with the `prefLabel` "shuttlecock" (used in badminton). Note that this concept has a number of `altLabel` predicates, one of which is indeed "bird".

Predicate	Object
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2004/02/skos/core#Concept
http://www.w3.org/2004/02/skos/core#prefLabel	"shuttlecock"
http://www.w3.org/2004/02/skos/core#altLabel	"bird"
http://www.w3.org/2004/02/skos/core#altLabel	"birdie"
http://www.w3.org/2004/02/skos/core#altLabel	"cock"
http://www.w3.org/2004/02/skos/core#altLabel	"shuttle"
http://www.w3.org/2004/02/skos/core#definition	"A shuttlecock is a high-drag projectile used in the sport of badminton. I

Figure 14. Kowari Query Result: Concept with altLabel Bird

These examples barely scratch the surface of navigating the micro "semantic web" we have created from this fairly simple spreadsheet. In the near future, we hope to modify the default Kowari interface to produce more user-friendly queries. For example, we might want to hide the URIs by mapping them to the objects that they represent so that the user is clicking only on English words.

10. Next Steps for SKOSaurus

We recognize the limitations of our pilot but, at the same time, we have achieved our proof-of-concept goals. Here is a partial list of features we are considering for implementation:

1. Normalize Definitions - The strongest model for assimilating data involves an approach that encourages contribution from multiple sources with contributors of varying degrees of skill. However, achieving a cohesive model through programmatic means without human intervention is unlikely in the near future, although the proper workflow can help us achieve the goal. A submission-and-approval process can be added which requires each Community of Interest (COI) to designate an owner who can promote entries from the candidate status to the approved status.
2. Edits of Existing Terms - Edits of an existing term would occur after a query. The query result screen could have a button such as [Edit Entry]. Once the button is clicked, the web form entry page could be loaded with the preexisting data. The user would have the opportunity to edit the fields and press [Submit]. At that point, processing would occur exactly like the current web form data entry.
3. Add Query across COIs.
4. Add historyNote element to each concept - to record CSV file upload date, contributor; update with modifications.
5. Examine Z39.19 in more detail since this may suggest other useful features we could incorporate.
6. If a database can output data in RDF or SKOS format, the SKOSaurus system could permit an upload or SOAP entry of the data without modification.
7. Migrate from an existing representation (database) to SKOS or RDF. At the time of this writing the new Oracle 10G Release 2 has exciting promise for supporting RDF using "Oracle Spatial". Oracle's Semantic Technologies Center describes this capability by saying "Oracle Spatial 10g introduces the industry's first open, scalable, secure and reliable RDF management platform. Based on a graph data model, RDF triples are persisted, indexed and queried, similar to other object-relational data types".
8. Support non-migration involving a separate and logically different backend that contains useful data, but not in our desired format. Since we can't expect everyone to provide data that we can enter directly into our system, we would consider other conversion processes. For example, ingestion of Microsoft Word tables would be a highly desirable feature.

9. Integrate Google define:, WikiPedia, and other generated search links.
10. Create XSLT and XSL-FO stylesheets for printing, probably with an interface to describe what subset of concepts to print.
11. Consider a central or federal data store for all concepts for multiple COIs and agencies with system management roles at COI and agency level, as well as across the entire data store.
12. Study the [NBII SKOS Thesaurus](http://nbii-thesaurus.ornl.gov/thesaurus/skosThesaurusSearch.jsp) [http://nbii-thesaurus.ornl.gov/thesaurus/skosThesaurusSearch.jsp] user interface and consider adding some similar features.
13. Add a Help button to the web entry form to explain all of the SKOS terminology. Similarly, add comments to the heading labels in the spreadsheet.
14. Enhance the web form so that repeatable fields (narrower, broader, related, altLabel, etc.) can be entered more than once.

11. Conclusion

Hopefully, we have demonstrated how an open source solution can be used to create an extremely useful application for the federal government (or for any COI). A web-based thesaurus created with thesaurus standards in mind, implemented in SKOS, and using proven web technology can hide most of the details for both the authoring and end user communities. That is, authors can simply define terms and indicate relationships in a natural way, using common applications such as spreadsheets. SKOSaurus can convert uploaded concept information into a SKOS/RDF form that greatly facilitates semantic discovery and manipulation of the terms.

Acknowledgements

Thanks to SAIC managers, Clive Carpi and Mary Ann Melosh, as well as U.S. Government thinkers and movers, Mike Daconta, Owen Ambur, Brand Niemann, Susan Turnbull, and Ira Grossman for their inspiration, interest in, and/or support of this pilot. Thanks also to Judith Newton (Ashton Computing and Management Services, LLC) for her insights in the early days of this project.

Bibliography

[Apache] *Apache HTTP Server Project*. Available at: <http://httpd.apache.org/> .

[CAF] *Chief Architects Forum*. Available at: <http://colab.cim3.net/cgi-bin/wiki.pl?ChiefArchitectsForum> .

[CGI] *CGI.pm - a Perl5 CGI Library*. Available at: <http://stein.cshl.org/WWW/CGI/> .

[CLS] *Concept-oriented with Links and Shared references Framework*. Available at: <http://www.ttt.org/clsframe/index.html> .

[DRM] *Federal Enterprise Architecture Data Reference Model*, September 2004. DRM 1.5 expected December 2005. Available at: <http://www.whitehouse.gov/omb/egov/a-5-drm.html> .

[DRM-WG] *Data Reference Model Working Group*. Available at: <http://colab.cim3.net/cgi-bin/wiki.pl?DataReferenceModel> .

- [EGov-Act] *E-Government Act of 2002*, December 2002. Available at: http://www.cio.gov/archive/e_gov_act_2002.pdf .
- [FEA] *Federal Enterprise Architecture*. Available at: <http://www.whitehouse.gov/omb/egov/a-1-fea.html> .
- [Glossary] *Glossary of terms relating to thesauri and other forms of structured vocabulary for information retrieval*; Willpower Information. Available at: <http://www.willpowerinfo.co.uk/glossary.htm> .
- [GlossXML] *Proposed XML Format for Glossaries; A recommendation for transporting dictionary information (GlossXML)*, 2002. Available at: <http://www.creativyst.com/Prod/Glossary/Doc/XMLOut.htm> .
- [IC] *U.S. Intelligence Community*. Available at: <http://www.intelligence.gov/1-members.shtml> .
- [IC-MWG] *Intelligence Community Metadata Working Group*.
- [ISO-2788] *ISO 2788*, 1986. Available at: <http://www.iso.org/iso/en/CombinedQueryResult.CombinedQueryResult?queryString=2788> .
- [ISO-1087] *ISO 1087*, 2000. Available at: <http://www.iso.org/iso/en/CombinedQueryResult.CombinedQueryResult?queryString=1087> .
- [ISO-704] *ISO 704*, 2000. Available at: <http://www.iso.org/iso/en/CombinedQueryResult.CombinedQueryResult?queryString=704> .
- [ISO-15836] *ISO 15836*, 2003. Available at: <http://www.niso.org/international/SC4/n515.pdf> .
- [Kowari] *Kowari Metastore*. Available at: <http://www.kowari.org/> .
- [MARTIF] *Machine-Readable Terminology Interchange Format [also known as ISO (FDIS) 12200]*. Available at: <http://coral.lili.uni-bielefeld.de/~ttrippel/terminology/node82.html> .
- [NBII] *NBII Thesaurus Web Service Prototype - Using SKOS* . Available at: <http://nbii-thesaurus.ornl.gov/thesaurus/skosThesaurusSearch.jsp> .
- [OLIF] *Open Lexicon Interchange Format*, April 2005. Available at: <http://www.olif.net/> .
- [Perl] *Perl.com*. Available at: <http://www.perl.com> .
- [PerlProg] *Programming Web Services with Perl*, December 2002; Randy J. Ray, Pavel Kulchenko. Available at: <http://www.oreilly.com/catalog/pwebserperl/> .
- [RDF] *Resource Description Framework*, February 2004. Available at: <http://www.w3.org/RDF/> .
- [SALT] *Standards-based Access to Multilingual Lexicons and Terminologies*. Available at: <http://www.loria.fr/projets/SALT/> .
- [SICoP] *Semantic (XML Web Services) Interoperability Community of Practice (SICoP)*. Available at: <http://web-services.gov/> .
- [SKOS] *Simple Knowledge Organisation System (home page)*. Available at: <http://www.w3.org/2004/02/skos/> .
- [SKOS-Core] *SKOS Core Guide*, May 2005. Available at: <http://www.w3.org/TR/swbp-skos-core-guide> .
- [SKOS-Intro] *Introducing SKOS*, June 2005. Available at: <http://www.xml.com/pub/a/2005/06/22/skos.html> .
- [SKOS-Vocab] *SKOS Core Vocabulary Specification*, May 2005. Available at: <http://www.w3.org/TR/swbp-skos-core-spec> .

[SKOS-Quick] *Quick Guide to Publishing a Thesaurus on the Semantic Web*, May 2005. Available at: <http://www.w3.org/TR/swbp-thesaurus-pubguide> .

[SOAP] *SOAP Version 1.2 Part 0*, June 24, 2003. Available at: <http://www.w3.org/TR/soap12-part0/> .

[Solaris] *Solaris 10*. Available at: <http://www.sun.com/software/solaris/> .

[VMware] *VMware - Virtual Infrastructure Software*. Available at: <http://www.vmware.com/> .

[WordNet] *WordNet, a lexical database for the English language*. Available at: <http://wordnet.princeton.edu/> .

[WXS] *XML Schema, Second Edition*, October 2004. Available at: <http://www.w3.org/XML/Schema> .

[XLT] *XML representation of Lexicons and Terminologies*. Available at: <http://www.ttt.org/oscar/xlt/dxlt.html> .

[XMLCoP] *XML Community of Practice*. Available at: <http://xml.gov/> .

[Z39.19] *XANSI/NISO Z39.19-2003, Guidelines for the Construction, Format, and Management of Monolingual Thesauri*, 2003; ISBN: 1-880124-04-1. Available at: <http://www.niso.org/standards/resources/Z39-19.pdf> .

Biography

Kenneth Sall

XML Data and Systems Analyst

[SAIC](http://www.saic.com/) [http://www.saic.com/]

Reston

Virginia

United States of America

<http://kensall.com>

Kenneth B. Sall is an SAIC XML Data and Systems Analyst. He previously served as XML Specialist on the GSA Integrated Acquisition Environment eGov Initiative. Ken created the XML section of [Web Developers Virtual Library](http://wdvl.internet.com/Authoring/Languages/XML/) [http://wdvl.internet.com/Authoring/Languages/XML/]. Addison-Wesley published his book, *XML Family of Specifications: A Practical Guide* [http://wdvl.internet.com/Authoring/Languages/XML/XMLFamily/], in June 2002. His [personal Web site](http://kensall.com) [http://kensall.com] contains several useful XML resources including his unique *Big Picture of the XML Family of Specifications*, an imagemap gateway to all major XML technical specifications which also indicates their maturity and depicts interrelationships. Ken has been an active participant in the Federal CIO Council's XML Community of Practice, the Semantic Interoperability Community of Practice, the FEA Data Reference Model (DRM) Working Group, and in the E-Forms for E-Gov Pilots.

Ronald Reck

Principle

[RRecktek LLC](http://iama.rrecktek.com/index.html) [http://iama.rrecktek.com/index.html]

Chantilly

Virginia

United States of America

[Ronald P. Reck](http://iama.rrecktek.com/rreck/rreck.doc) [http://iama.rrecktek.com/rreck/rreck.doc] was raised and educated in the Detroit Metropolitan area and on occasion, has enough time to miss the friends and culture of the place he still calls home. He is formally trained in theoretical syntax and remains fascinated by language and what it reveals about being human. A passion for linguistics and intensity with computers afford him gainful employment using Perl, XML, and Semantic Web technologies running, of course under *nix. He prides himself on developing scalable, open source architectural strategies for difficult problems. He resides near our nation's capital with his lovely wife Olga and two cats.