Electronic Commerce Connection, Inc.

XML Taxonomies

Prepared for:

Betty Harvey

harvey@eccnet.com

XML Taxonomies

Betty Harvey
Electronic Commerce Connection, Inc.
harvey@eccnet.com
410-787-9200

Table of Contents

Introduction

This presentation will discuss various issues related to developing a taxonomy in XML. It will discuss some of the current efforts to develop taxonomic data for biological data, as well as provide insight into taxonomic develop outside the realm of traditional taxonomic data.

Slide 1: XML Taxonomies

Betty Harvey
Electronic Commerce Connection, Inc.
harvey@eccnet.com

Section 1: What is a Taxonomy

Introduction

This section will talk only about the definition and differences between taxonomy, classification and ontology. These words are used interchangeably. However, there are nuances and differences between terminology.

Slide 2: What is a Taxonomy

Slide 3: Taxonomic Vocabulary

- Taxonomyorderly classification of information according to their presumed natural relationships.

  NOTE: The word Taxonomy originated from the Greek words
  - taxis meaning arrangement or division
  - and nomos meaning law
- Taxon A taxonomic group or entity
- TaxaPlural of Taxon
- Classification systematic arrangement in groups or categories according to established criteria
- Ontologya particular theory about the nature of being or the kinds of existents

Instructors Note:

During this slide I want to just bring up the 'raw' dictionary definition for each one of these terms. I will discuss the nuances later on in the talk.

After watching "My Big Fat Greek Wedding" I thought it was appropriate that the word 'taxonomy' is related to Greek.

Slide 4: Taxonomies are Universal

People use both formal and informal taxonomies everyday.

- TV Guide

Human beings develop personal taxonomies everyday.

- Grocery lists/Recipes
- Todo lists
- Contact lists

Instructors Note:

My goal with this slide is to show that we really use taxonomies in everyday life. Most individuals don't understand or recognize that they are developing taxonomies, they believe they are 'making lists'.

Slide 5: Taxonomy vs. Classification vs. Ontology

Taxonomy, classification, and ontology are very often used interchangeably. However, there are subtle differences.

- Taxonomy: a hierarchical arrangement of topics that imposes topical structure on information in a specific body of knowledge.
- Classification: the process of dividing objects or concepts into logically hierarchical classes, subclasses, and sub-subclasses based on the characteristics (attributes) they have in common and those that distinguish them. Note, this is the model upon which individual taxonomies may be built.
- Ontology: An ontology is a knowledge representation system which presents the key concepts and relationships relevant to a body of knowledge

NOTE: *Definition posted on ebXML Registry/Repository listserve by Nicholas Berry, Librarian, Boeing XML Registry*

Instructors Note:

One of the problems today is the word 'taxonomy' is being used to define anything. Classic information management systems are being called taxonomies. The words Taxonomy, Classification and Ontology are often used interchangeably and the differences are subtle but important.

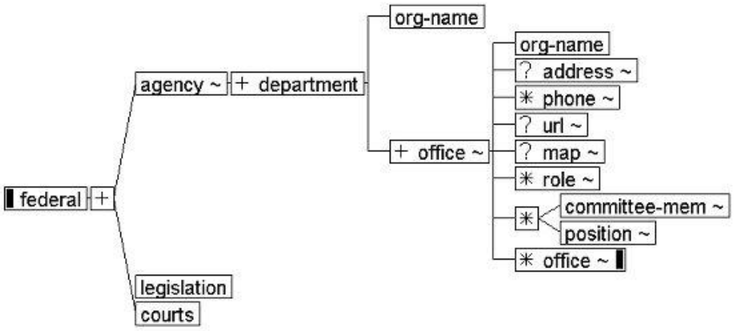
I really like Nicholas' definitions for this terminology, especially in an XML environment. In the next few slides I will attempt to discern how this terminology can be related to an XML application.

Slide 6: XML Taxonomy

Taxonomy - usually hierarchical arrangement of topics that imposes topical structure on information in a specific body of knowledge.

- DTD
- Schema (W3C, Relax)

Model of an XML Taxonomy



Instructors Note:

This slide shows a hierarchical model for information which models the federal government hierarchy. It is incomplete but shows how the relationship of information modeled.The next slide will show how this model can be tagged at the office level. At this point, I can walk through the entire schema to show the relationships.

Slide 7: XML Classification

Classification - Inclusion of data into the information model or schema.

Classification based on XML Schema

```
<?xml version="1.0"?>
<?xml-stylesheet href="position.xsl" type="text/xsl"?>
<!DOCTYPE office SYSTEM "government-org.dtd">
<office>
  <org-name>National Archives</org-name>
  <phone type="phone">
  <areacode>202</areacode>
  <number>555-1212</number>
  </phone>
  <phone type="fax">
  <areacode>202</areacode>
  <number>555-1213</number>
  </phone>
  <phone type="other" other="TTY">
  <areacode>202</areacode>
  <number>5551214</number>
  </phone>
  <position>
  <position-title>Chief Scientist</position-title>
  <person>
  <name>
    <fname>Benjamin</fname>
    <lname>Franklin</lname>
    <suffix>Ph.D</suffix>
  </name>
  <email>benjamin.franklin@heaven.org</email>
  <url>http://bensguide.gpo.gov/</url>
  <picture><graphic name="../graphics/franklin.gif"/>
  </picture>
  <bio>
    <p> Franklin was born in 1706 at Boston. He was the tenth son of a soap
    and candlemaker. He received some formal education but was principally
    self-taught. After serving an apprenticeship to his father between the ages of
    10 and 12, he went to work for his half-brother James, a printer. In 1721 the
    latter founded the New England Courant, the fourth newspaper in the colonies.
    Benjamin secretly contributed 14 essays to it, his first published writings.
    </p>
    <skills-inventory>
    <skill>
    <skill-name>Printer</skill-name>
    <proficiency>Master</proficiency>
    <certification>None</certification>
    </skill>
    <skill>
    <skill-name>Scientist</skill-name>
    <proficiency>Electricity</proficiency>
    <certification>National Kite Award</certification>
    </skill>
    <skill>
    <skill-name>Inventor</skill-name>
    <proficiency>Bifocals, Lightning Rods, Model Basins, Watertight
```

```
        Bulkheads, Odometer</proficiency>
      <certification>None</certification>
      </skill>
      <skill>
      <skill-name>Statesmen</skill-name>
      <proficiency>Expert</proficiency>
      <certification>Signed the Declaration of Independence and the U.S.
        Constitution</certification>
      </skill>
      <skill>
      <skill-name>Musician</skill-name>
      <proficiency>Premier</proficiency>
      <certification>Violin, Harp, Guitar and Armonica</certification>
      </skill>
      </skills-inventory>
    </bio>
    </person>
    </position>
    <position>
    <position-title>Chief Inventor
    </position-title>
    <person>
    <name>
      <fname>Thomas</fname>
      <mname>Alva</mname>
      <lname>Edison</lname>
    </name>
    <picture><graphic name="../graphics/edison.jpg"/>
    </picture>
    <bio>
      <p>Thomas Edison was not born into poverty in a backwater mid-western
      frontier town. Actually, he was born (on Feb. 11, 1847) to middle-class parents
      in the bustling port of Milan, Ohio, a community that - next to Odessa, Russia
      - was the largest wheat shipping center in the world. In 1854, his family moved
      to Port Huron, Michigan, which ultimately surpassed the preeminence of both
      Milan and Odessa....</p>
      <skills-inventory>
      <skill>
      <skill-name>Newsboy</skill-name>
      <proficiency>Mediocre</proficiency>
      <certification>Grand Trunk Railroad</certification>
      </skill>
      <skill>
      <skill-name>Telegraph</skill-name>
      <proficiency>Expert</proficiency>
      <certification>No Certification</certification>
      </skill>
      <skill>
      <skill-name>Electricity</skill-name>
      <proficiency>Expert</proficiency>
      <certification>Invented the lightbulb</certification>
      </skill>
      </skills-inventory>
    </bio>
    </person>
    </position>
  </office>
```

Link to Styled Example locally or Link to Styled Example on the Internet

Instructors Note:

In this example, I want to show how the data can be incorporated into the model. I also want to show how the data can be viewed within a browser. Sometimes it is hard to visualize how and XML stream can be visualized by the 'common-person'.

If you are viewing this presentation in a PDF file, then you should be able to link to the styled example via the Link to Styled Example on the Internet.

Slide 8: XML Ontology

Ontology - Knowledge management systems developed around the Taxonomy/Classification

- Web Servers
- Retrieval systems
- Distributed Systems
- Topic Maps
- etc.

Instructors Note:

Ontology and Taxonomy are both used interchangeably. I see ontologies being the knowledge management systems that use taxonomies. Taxonomies become really useful with a knowledge management system. Topic maps have a definite role in creation of ontologies.

Section 2: Developing an XML-based Taxonomy

Introduction

This section will outline the requirements for developing an XML-based taxonomy. The section will deal with 'generic' taxonomies and not any specific taxonomies. Several examples will be given.

I have been involved in developing several taxonomic models for various types of content. One of the areas that I want to stress in this section is that the information analysis phase is the most important phase of the project. If the information analysis isn't complete, then the taxonomy will always be flawed.

More organizations want to short-cut the analysis phase of the project. It costs organizations more in the long run to revise broken models than to do the analysis and get them right the first time.

Slide 9: Developing an XML-based Taxonomy

Slide 10: Resources Required for Developing an XML Taxonomy

A taxonomy requires many types of individuals in order to be successful.

- Subject Matter Experts (SME)
- Domain Experts
- Information Technologists
- Resource Users

Instructors Note:

It is important that all the individuals involved in creating and using the information be involved in developing the taxonomy. An IT or XML expert can develop an elegant taxonomy solution but if it isn't used then the taxonomy is useless.

Slide 11: Understanding the Requirements

It is important to have an understand:

- **The relationship of the information**
- **Who will use the information**
- **How they will use the information**
- **Granularity of information**
- *How the information will be searched*
- *How the information will be retrieved*

Instructors Note:

Often times, going into an XML project the entire project is not solidified. The first four bullets are essential to know before you can begin. However, the last two bullets are often unknown. In some cases you have to make the best guess about how the information will be searched and retrieved.

Slide 12: Information Analysis

The most important process in developing the taxonomy or an XML Schema

- Information Analysis is a modeling exercise
- May be bottom-up or top-down Hierarchical Models

Instructors Note:

The information analysis is the most important phase of the taxonomy project. If the information model is wrong then the taxonomy will not be very useful or require revision.

Slide 13: Develop a Common XML Architecture

Decide on an XML Architecture

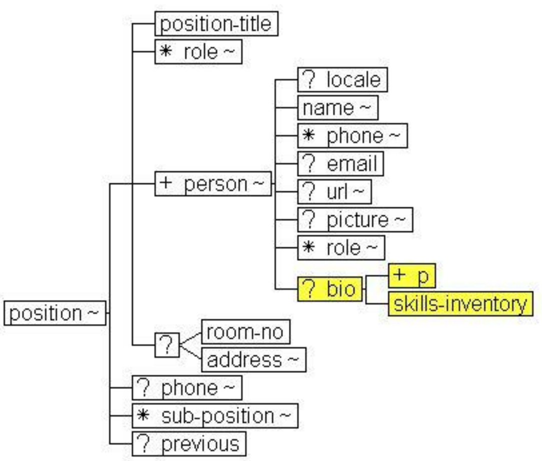
- Naming Conventions
- Abbreviations
- Elements vs. attribute philosophy
- Incorporated standards
- Generated Text
- Other information depending on the project

Instructors Note:

When developing a common architecture you need to look at all the items above. You will also have an understanding of any other information systems that may need to be integrated or interact with the taxonomy.

Slide 14: Consider Modularization

Modularized DTD



Slide 15: Modular Example

```
<position xmlns:xlink="http://www.w3.org/1999/xlink">
  <position-title>Chief Scientist</position-title>
  <person>
  <name>
    <fname>Benjamin</fname>
    <lname>Franklin</lname>
    <suffix>Ph.D</suffix>
  </name>
  <email>benjamin.franklin@heaven.org</email>
  <url>http://bensguide.gpo.gov/</url>
  <picture><graphic name="../graphics/franklin.gif"/>
  </picture>
  <bio xlink:href="http://www.eccnet.com/~bfranklin/bio.xml">
  </person>
  </position>
```

- Provides dynamic information.
- The information stakeholder authors the information
- Allows distributed processing
- Better security

Slide 16: Good Taxonomy Practices

- Use the vocabulary of the domain space
- Develop appropriate relationships between information objects
- Understand the requirements for use of the taxonomy
- Try to accommodate the current and future use of the taxonomy

Instructors Note:

Use the vocabulary that the SME's and domain users use. IT types tend to name information in their own terminology. If the terminology is not appropriate for the subject then the taxonomy will not be used by domain users.

Slide 17: XML - A Very Good Format for Taxonomies

- Taxonomies are highly structured data
- Relationship building required for taxonomies
- XML strength is structured data
- Taxonomy information is usually created by multiple sources
- Good for distributed information

Section 3: Taxonomy and Relationships

Slide 18: Taxonomy and Relationships

Slide 19: Taxonomy is Highly Structured Data

- Classification Scheme
- Relationship Building
- Usually hierarchical (*Talk more about this in the next slide)*
- Relationships should be consistent

Slide 20: Taxonomies Contain Anomalies

- Rules are made to be broken.
- Every taxonomy has exceptions to the rule.
- XML modelings must provide representation for when rules are broken.

Instructors Note:

When building taxonomies, you must take into account where the rules are broken or modified depending upon the situation. If you are creating a new taxonomy using new data, then it is much easier to develop strict taxonomy models. However, when developing taxonomies from historical data the data is not always clean.

Slide 21: Genealogy

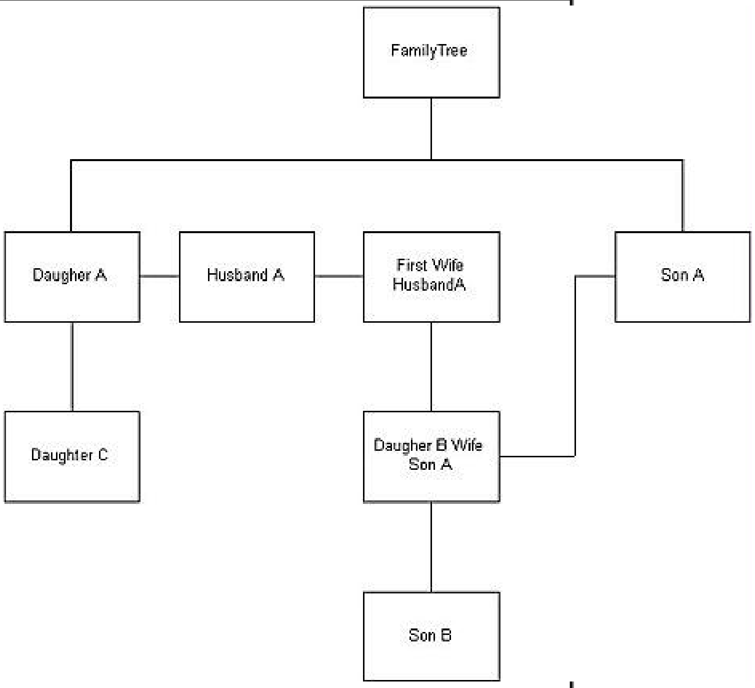
Genealogy is a classic taxonomy. A family tree should always branch and shouldn't connect - right? -- wrong?

- **Types of Relationships**

  Relationship of Daughter B to Son A
  - Niece (Step)
  - Wife

  Relationship of Daughter B to Son B
  - Mother
  - Cousin (Step)

  Relationship of Son B to Husband A
  - Grandfather
  - Uncle



Instructors Note:

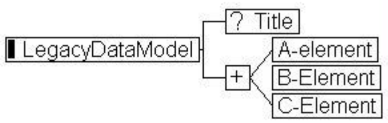
This diagram shows that even a classic hierarchical taxonomy like a family tree can have anomalies. Notice how the hierarchy is now cyclical. When developing taxonomies a understanding of the possibilities of the information being broken is a possibility. Accommodations need to be made in the XML to accommodate these anomalies.

Slide 22: New vs. Legacy Taxa

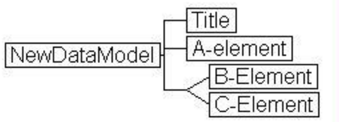
Legacy taxa require looser models.

- Accommodate all historical data which may be inconsistent
- New data can be developed in a more consistent manner
    - Consider a tighter schema where the data can be validated against the looser model for legacy data.

Legacy Data Model



New Data Model



Instructors Note:

In this slide I want to show how data models for legacy and new data can be harmonized. The legacy data model is a looser model. If you look at the *Legacy Data Model*, a title is optional and the A-element, B-Element and C-Element can be allowed in any order in as many times are the legacy data has been created.

However, in the *New Data Model* we want to ensure conformance to the required data model. In this case, a Title is required followed by a required A-Element followed by either a B-Element or C-Element. New data can be parsed against both the LegacyDataModel schema or the NewDataModel schema.

Section 4: Biologia Centrali-Americana (BCA)

Slide 23: Biologia Centrali-Americana (BCA)

Slide 24: Biologia Centrali-Americana (BCA)

A joint project of :

- Smithsonian Institution (National Museum of Natural History, Smithsonian Institution Libraries, and Smithsonian Tropical Research Institute)
- Natural History Museum (London)
- Missouri Botanical Garden
- National Commission for the Knowledge and Use of Biodiversity, Mexico (CONABIO)
- Instituto Nacional de Biodiversidad, Costa Rica (INBio)
- American Museum of Natural History
- Harvard University (Museum of Comparative Zoology)
- Royal Botanic Gardens, Kew
- Museo Entomologico de Leon, Nicaragua
- Global Biodiversity Information Facility

Slide 25: BCA

- 63 volumes.
- Out-of-print.
- Fundamental resource for the study of Central American flora and fauna.
- Compiled from scientific surveys and explorations during the 19th and 20th centuries.

Slide 26: BCA Goals

Multi-phase project to create tools and resources for biodiversity studies.

Phases:

- Information Analysis
- Create a global schema for taxonomic literature.
- Scanning, rekeying and coding 58 of the 63 BCA volumes
- XML database
- Linking to other biological datasets

Instructors Note:

Information available at: http://www.sil.si.edu/bcaprojectProject Overview Document (2/5/2003) Smithsonian Institution Libraries (SIL) has received a grant to complete the first stage of a multi-phase project that will create a model set of electronic tools and resources for biodiversity studies. The first phase consists of scanning, rekeying, and coding in eXtensible Markup Language (XML) the contents of 58 volumes of the out-of-print Biologia Centrali-Americana, a fundamental resource for the study of Central American flora and fauna compiled from scientific surveys and explorations conducted during the late 19th and early 20th centuries.

Many of the period's eminent biologists contributed specimens and descriptions, and the accompanying illustrated plates are often the only images of Central American biota in existence. The coded text will be mapped to a database and linked to other vital biological datasets, including the National Museum of Natural History (NMNH) collections information system (the NMNH Multimedia Catalogue System), so scientists world-wide can study specimens in the context of the published scientific record. With staff of NMNH providing scientific guidance, the project will serve as a model for several other major bioinformatics projects pursued by the world's leading biological repositories. Together these will speed the pace of scientific

	investigation into the nature of our rapidly changing biological environment.

Slide 27: A Few Challenges in Developing BCA Taxonomy

- Expected to be a "Global Schema"
- Hierarchical model (recursive vs. explicit)
- Page Fidelity
- Integration of Metdata
- Integration of TEI-Lite

Slide 28: Global Schema

- Tried to make a global schema for taxonomic literature
- Only had access to BCA
  - Relying heavily on subject matter experts knowledge of other literature
  - Model must be very loose to accommodate the unknown
  - BCA will be the first test of schema
- Schema has to accommodate both scientific and library services requirements

Slide 29: Biology Hierarchical Classifications

Kingdom	Phylum	Division
Subphylum	Subdivision	Superclass
Class	Subclass	Superorder
Order	Suborder	Superfamily
Family	Subfamily	Tribe
Subtribe	Group	Genus
Subgenus	Section	Subsection
Series	Subseries	Species
Subspecies	Variety	Subvariety
Form	Subform	

Instructors Note:

The elements above show the order of classification of species. BCA, as well as other taxonomic literature do not always conform to the hierarchy above.

Slide 30: Ordered vs. Recursive Model

Three options for developing top-level hierarchy:

(1) Explicit Named Hierarchical Model

```
<Kingdom>
      <Phylum>
            <Division>
               ...
            </Division>
      </Phylum>
</Kingdom>
```

(2) Recursive Model

```
<TaxonTreatment RankDesignation="Kingdom">
      <TaxonTreatment RankDesignation="Phylum">
           <TaxonTreatment RankDesignation="Division">
               ...
           </TaxonTreatment>
      </TaxonTreatment>
</TaxonTreatment>
```

(3) Explicit Hierarchical Model Using Number Algorithm

```
<TaxonTreatment1>
     <TaxonTreatment2>
          <TaxonTreatment3>
             ...
          </TaxonTreatment3>
      </TaxonTreatment2>
<TaxonTreatment1>
```

Instructors Note:

We had three options:

- Create an explicit hierarchy by naming the top level and lower levels For example a <class> could contain one or more <order>, which could themselves contain one or more <suborder>, etc.
- Use a recursive structure. In recursion, a structure would be allowed to contain itself, thus a <TaxonTreatment> could contain one or more <TaxonTreatment>s, which could

themselves could themselves contain <TaxonTreatment>s, etc. The type of structure can be identified using an attribute value, e.g. <TaxonTreatment type="class"

- Create an explicit hierarchy by numbering the top level as "1" and subsequent levels as "2", "3", etc. (Thus a <TaxonTreatment1> could contain one or more <TaxonTreatment2>s, which could themselves contain one or more <TaxonTreatment3>s, etc.).

There are advantages and disadvantages to each approach. Explicit hierarchy is easier to process and is readily understandable. In newly created/authored data defining explicit hierarchy is the appropriate mechanism to ensure consistency of authored information.

However, legacy or previously published data does not always conform to the established hierarchy because it has been created without the advantages of constrained rules. The rules for creating data are usually implied but there wasn't an application to constrain the defined rules. The BCA project and other taxonomic information has been authored over 100 years. There will be inconsistencies in the authoring of the information. Until actual conversion of the data begins all the variations in the data will not be known. The taxonomic schema will attempt to accommodate known inconsistencies.

After much discussion it was determined that the upper-level hierarchical information will be a recursive structure with an attribute that identifies the structure level.

A decision was made to using a recursive model.

Slide 31: Page Fidelity

Requirement for Page Fidelity:

- Other literature reference page numbers
- Important to be able to search page and volume and get to the right location.

Solution:

- PageBreak Element
- Appropriate Attributes

```
<PageBreak Pagenumber="1"
         LinkToMetadata="danais-page1"
         LinkToOriginalPage="danais-page1"
         PageHeader="RHOPALOCERA."/>
```

Instructors Note:

BCA is an historic taxonomic publication that is used and has been used for the past 100+ years. Many publications cite the BCA by page number and species. Therefore, it was essential that a 'hook' be available to reference the exact printed page. Also, the scanned images of the pages will be available to be viewed. This information will be done with the <PageBreak> element.

Slide 32: Integration of Metadata

- Very little metadata included in BCA
- Metadata already available externally
- Should the metadata be incorporated internally in data or continue to be maintained externally.
- Metadata standard has not been solidified
  - MARC
  - MODS
  - METS
  - Dublin Core
  - Access to Biological Collection Data (ABCD)
  - Federal Geographic Data Committee (FGDC)

Decided to have the metadata records external to data and the ability to link to metadata internally. The <PageBreak> element will be used for external metadata linking.

Instructors Note:

Metadata always becomes an issue in any XML publishing project.. There are so many metadata standards available for use it is sometimes a difficult decision which metadata standard to use. These decisions are best left to the librarian or the publisher with advice from the IT staff.

For BCA, a decision was made to keep the metadata separate from the literature and provide a mechanism for the user to link to the appropriate metadata. The metadata may ultimately be used to aid the user to search and navigate through the data.

Slide 33: Integration of TEI

TEI was mandated to be used for front and back matter for volumes and individual taxonomic literature.

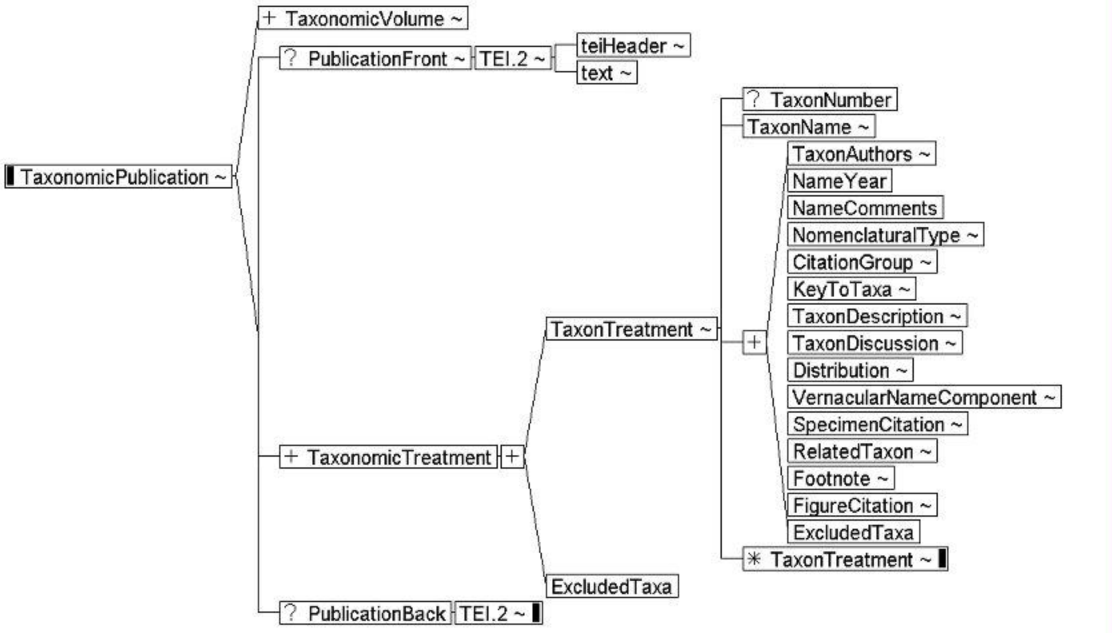
Issues and/or Problems

- Taxonomic/Scientific Models Camelcase - TEI lowercase
- Full Names - TEI uses mostly abbreviation, i.e., (pb = page break)
- Bibliographic models do not conform to requirement for taxonomic citations

Slide 34: Integration of TEI Solution

TEI would only be used for front and back matter:

Taxonomic Publication Model



Slide 35: Relationship/Linking Attributes

- Unknown information management system
- Need to be able to maintain context of relationship.
- ParentNode The unique identifier of the parent element.
- SiblingNodeHigher The unique identifier of a preceding sibling element.
- SiblingNodeNext The unique identifier of the following sibling element.

Slide 36: BCA Example

Styled TaxonomicPublication Example

Raw TaxonomicPublication Data

Section 5: About the Presenter

Introduction

ECC, Inc. is a small woman-owned business. I started ECC, Inc. in 1995. I have participated with many Government and commercial enterprises in planning and executing their migration to structured information. Over the past nine years I have developed many SGML and XML solutions for a wide-range of clients.

Prior to starting ECC, Inc., I worked in Scientific and Engineering Computing at David Taylor Model Basin, Navy Systems Weapon Center. While at David Taylor I participated in the development of U.S. DoD CALS standards, including IETMs, SGML and Internet protocols. In 1994, I was awarded "Employee of the Year, Engineer/Scientist". In 1994 I resurrected the Washington Area SGML Users group. I am still acting as the coordinator of the current Washington, DC Area XML Users Group.

I have written many articles for technology and e-business magazines. I am also one of the co-authors of "Professional ebXML Foundations" published by WROX.

I have taught many XML related courses both in the U.S. and in Europe for both government and private industry. The most exotic place I have taught is Kiruna, Sweden (128 kilometers above the arctic circle and original home of the "Ice Hotel").

Slide 37: Contact Information

Betty Harvey
Electronic Commerce Connection, Inc.
harvey@eccnet.com