Selecting a Taxonomy Management Tool

Wendi Pohs

InfoClear Consulting

#SLATaxo

InfoClear Consulting – What do we do?

- Content Analytics Strategy and Implementation, including:
  - Taxonomy/Ontology development and maintenance
  - Semantic technology and tool selection
  - Metadata tagging strategy
  - Search engine integration and relevance tuning
  - Application design for unstructured data
  - Graph database design

What we'll discuss today

- Overview of the basic, high-level considerations
- A deep dive into how taxonomy management software was selected at IBM
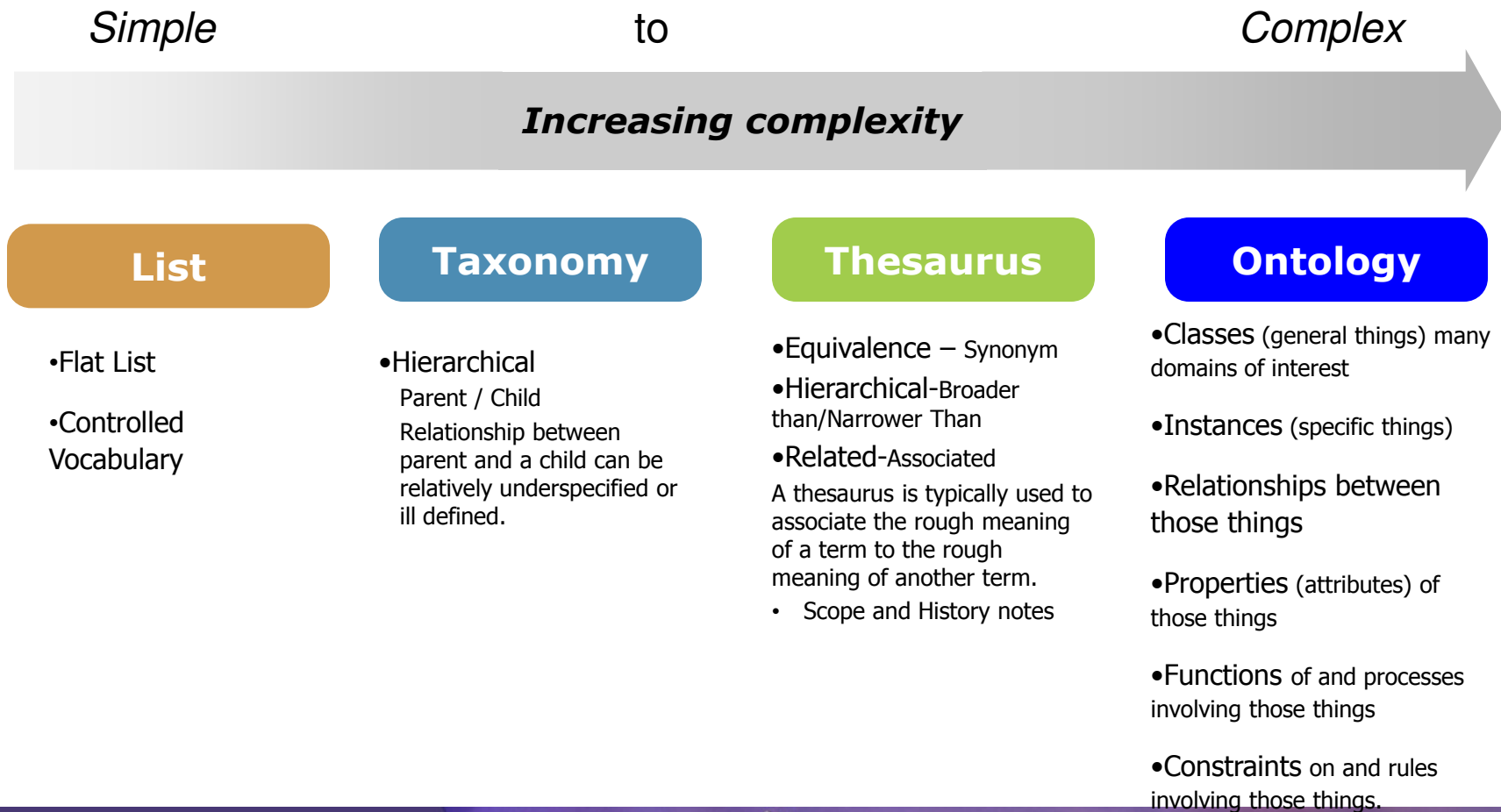- Questions and discussion

A quick recap: The basic components

- **Taxonomy** (from the Greek: taxis (order) or a classification scheme)
  - A dynamic list of words and phrases that are used to describe what content is about

- Taxonomies include:
  - **Facets** – A named group of similar words and phrases, used to provide input to metadata on the backend, and for different aspects of search
  - **Relationship Types** – Named links between words and phrases in facets
  - **Attributes** – Properties of terms
    - Scope notes, dates
  - **Entities** – People, Places and Things, often proper nouns that are easily identifiable within content
- **Metadata** – Named fields that are associated with content, often populated by the words or phrases in the taxonomy
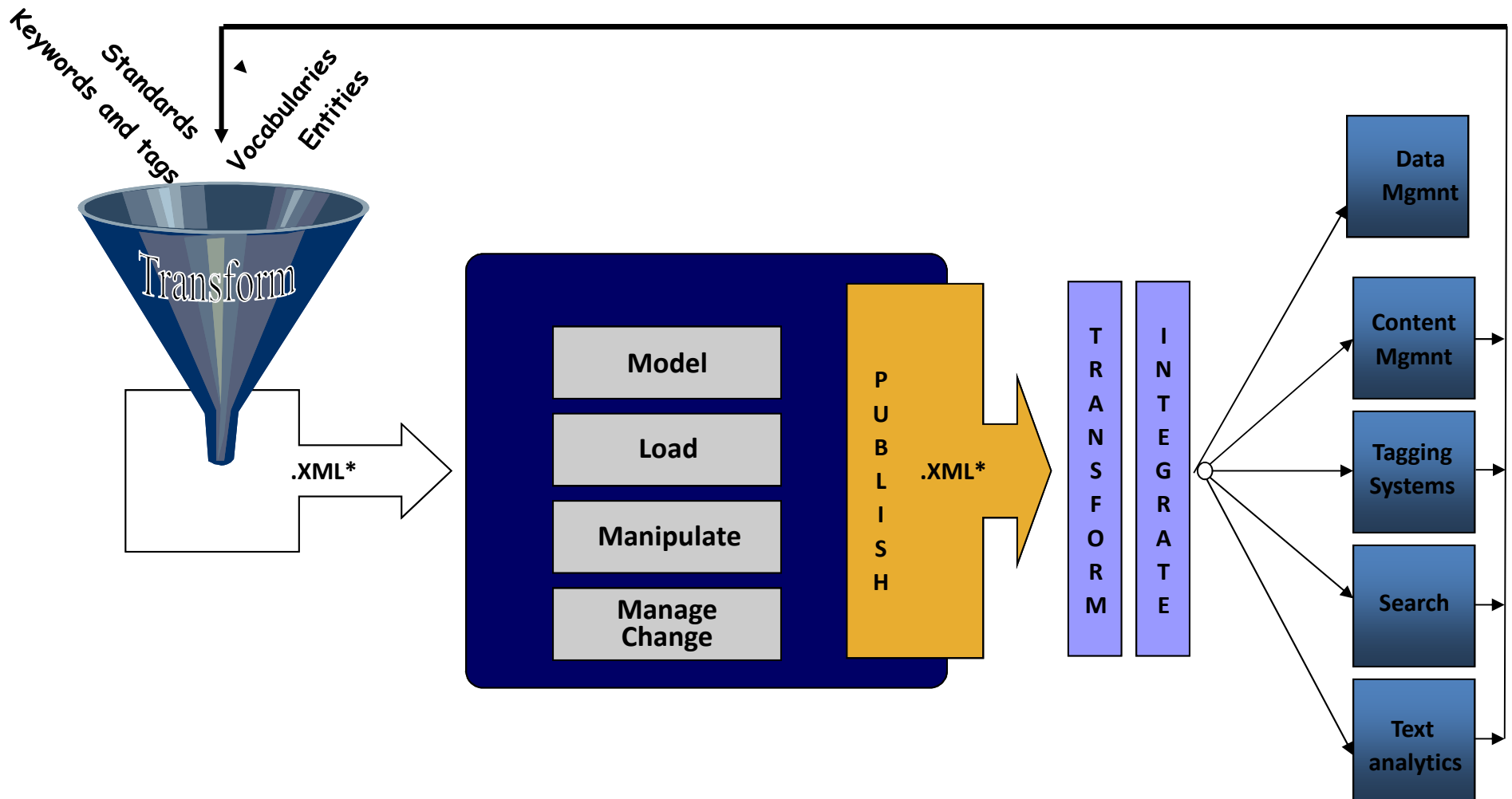
The Semantic Spectrum: What do you need to manage?

Simple　　　　　　　　　　　to　　　　　　　　　　　*Complex*

Increasing complexity

List	Taxonomy	Thesaurus	Ontology

List
- Flat List
- Controlled Vocabulary

Taxonomy
- Hierarchical
  Parent / Child
  Relationship between parent and a child can be relatively underspecified or ill defined.

Thesaurus
- Equivalence – Synonym
- Hierarchical-Broader than/Narrower Than
- Related-Associated
  A thesaurus is typically used to associate the rough meaning of a term to the rough meaning of another term.
- Scope and History notes

Ontology
- Classes (general things) many domains of interest
- Instances (specific things)
- Relationships between those things
- Properties (attributes) of those things
- Functions of and processes involving those things
- Constraints on and rules involving those things.

How are taxonomies used?

Keywords and tags
Standards
Vocabularies
Entities

Transform

.XML*

Model

Load

Manipulate

Manage Change

PUBLISH

.XML*

TRANSFORM

INTEGRATE

Data Mgmnt

Content Mgmnt

Tagging Systems

Search

Text analytics

As reference sources for other processes

- **Allowed Values Lists for Metadata** – A collected list of metadata values, most often used in Content Management systems

- **Tagging systems –** Input to manual and automated processes for applying metadata to content
  - **Auto-Categorization** – A set of techniques, primarily automated, used for grouping similar content together

- **Entity extraction** **–** Identifying and extracting specific terms and phrases from content, either from an Authority List, or based on known patterns in text

- **Transaction systems –** Support integration and mapping of similar metadata from different systems

- **Web sites –** Support site navigation and faceted search

Why buy a taxonomy management tool?

- Your taxonomy spreadsheet has become too cumbersome
    - No real tree view
    - Hard to visualize related terms
    - Hard when more than one person maintains the taxonomy
    - You need better hooks for integration

How can you justify your purchase?

Your purchase justification is most often based on consuming applications, and not on the taxonomy tool itself, so you need to understand how you will use it

– For auto-categorization?

– As part of a content management system?

– For Web site navigation?

– To support mapping between terms in legacy systems?
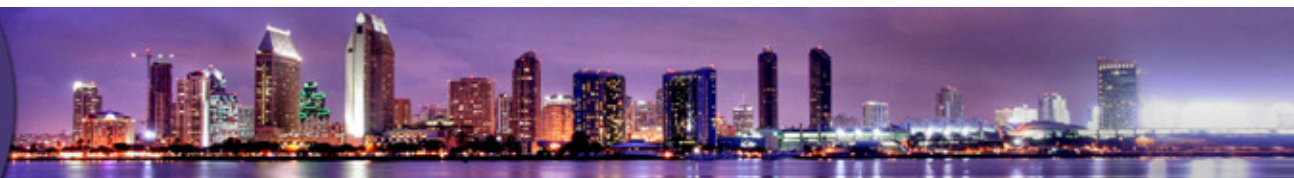
– For records or data management?

And you need to understand who will use the tool

- Are you working alone or will you be collaborating with others?
- Do you need to support other languages?
- Will you need IT support for installation and/or integration with other systems?
- Will this become an Enterprise tool?

What do the tools offer?

1. Data modeling
   - Building your taxonomy
     - Facets, mappings, relationship types, attributes
- Considerations
  - Do you need to manage a standard thesaurus with BT/NT relationship types?
  - Do you need other named relationships?
  - How does the tool handle synonyms?
  - How complex are your attributes?
  - How does the tool handle polyhierarchy?
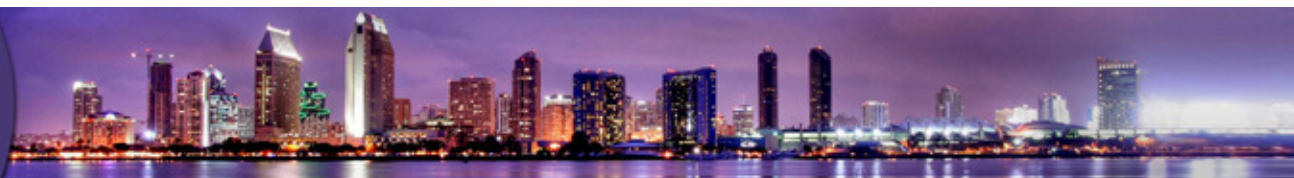
What do the tools offer?

2. Editing
   - Creating, renaming, merging and deleting terms, promoting and demoting terms within hierarchies, mapping terms
   - Managing relationships and attributes

- Considerations
  - What is your most common editing task? Does the tool easily support it?
  - Can you change many terms at once?
  - How easy is it to find the terms you need to edit?

What do the tools offer?

3. Import/Export functionality
   – To and from lists, spreadsheets, XML and other formats

- Considerations
   – Are the taxonomy standards, like SKOS or ZTHES, supported by the tool?
   – How robust is the spreadsheet import functionality?
   – What formats do your consuming applications need?

What do the tools offer?

4. Workflow
   - Draft => Approved/Published
   - Change notification
   - Email support

- Considerations
   - These vary widely between tools, so think about what you really need, for example:
     - Clear indication of term status if your taxonomy has many editors
     - Clear indication of when a term has been approved or published

What do the tools offer?

5. Integration interfaces
   - APIs
   - XML outputs
- Considerations
  - Understand the tools your developers use
    - Open source, Java, or C# tools?
  - Are there built-in connectors or integration components?
  - Does the tool operate on your platform or server?
    - Windows, Linux, Mac?
  - Include IT in your purchase decisions

What do the tools offer?

6. Advanced features, for example:
    - Visualization of terms and relationships
    - Built-in auto-classification tools
    - Built-in term mapping tools
    - Robust versioning and archiving support
    - Multi-lingual support
    - Text analytics
- Considerations
    - Visualization tools are important if you need to sell your taxonomy to management
    - Auto-classification tools can require a slightly different design methodology.
        - Understand the primary purpose of the tool
    - Versioning and archiving are not givens
    - Your environment and needs are unique

The intangibles

- Company viability and support
  - You can often judge responsiveness by the technical support you receive as you evaluate the software
- Take the time to get customer references
- Considerations
  - Taxonomy management is often an adjunct function when it is part of other tools.
  - Smaller companies can be more agile, but new features are often driven by the loudest customer voice
  - Consider tools that are easy to integrate with the related technologies that you use

General product considerations

- Commercial products
  - Suites of tools
  - Formal support
  - Formal roadmaps and development schedules
  - Customization is almost always necessary
  - Taxonomy management modules are sometimes created to support other, primary applications
- Open source products
  - Most flexibility for homegrown solutions
  - Crowd-sourced, but large user community

Questions? Need more info about specific tools?

wpohs@infoclearonline.com

Or visit our Website at:

http://www.infoclearonline.com