# Methods for the Automatic Construction of Topic Maps

**Steve Pepper, CEO, Ontopia**
**Convenor SC34/WG3, Editor XML Topic Maps**
pepper@ontopia.net

## Road Map

- **Overview of the issues involved**
  - Topic map basics
  - Tasks involved in creating topic maps
  - Sources of topic map data
  - Extraction techniques
- **In-depth examination of two cases**
  - Generating topic maps from XML structured data
  - Using Natural Language Processing on unstructured data
- **Application examples**
  - NLP with SemanText
  - Topic map erotica with the Ontopia MapMaker

## Topic map basics

- **Basic constructs:**
  - Topics: representing “subjects of discourse”
  - Associations: representing relationships between subjects
  - Occurrences: information relevant to a given subject
- **Types**
  - Topic types: classes of topics
  - Association types: classes of associations
  - Occurrence types: classes of occurrences
- **Types are also topics**
  - the “ontological” or “typing” topics

*Of course, there is more, but this is sufficient for our present purpose...*




## Examples

- **Topics:**
  - Ontopia, ISOGEN, Pepper, Freese
- **Associations:**
  - Pepper *is employed by* Ontopia
  - ISOGEN *is a partner of* Ontopia
- **Occurrences:**
  - Ontopia’s web site, Pepper’s bio, Freese's portrait
- **Typing topics:**
  - **topic types:** company, person
  - **association types:** employedBy, partnerOf
  - **occurrence types:** web site, bio, portrait

## Aspects of topic map creation

- **Identifying typing topics**
  - i.e., the topics that define classes (or types) of topics, associations and occurrences
  - e.g., “person”, “company”, “employedBy”, “partnerOf”, “web site”, “bio”
- **Identifying individual topics**
  - e.g., “Ontopia”, “ISOGEN”, “Pepper”, “Freese”
- **Identifying individual associations**
  - e.g., “Pepper *is employed by* Ontopia”
- **Identifying individual occurrences**
  - e.g., a bio of Pepper, a photo of Freese, an org. chart of ISOGEN
- *These tasks are significantly different in their complexity and the frequency with which they must be performed*




## A daunting task?

- **In one sense, yes, since human intellectual effort is required**
- **On the other hand, this is something humans do all the time, and much of it even gets recorded:**
  - data analysis leads to schemas and DTDs
  - classification leads to card catalogs, etc.
- **However, this effort is usually expended for a single purpose and is difficult to leverage for other purposes**
- **With topic maps, our goal is**
  - maximum leverage of whatever sources already exist
  - preservation of intellectual effort for future reuse

## Sources of topic map data

- **Structured knowledge**
- **Document metadata**
- **Structured document content**
- **Unstructured document content**
- **Information systems**
- **Knowledge in people's heads**




## Structured knowledge

- **Ontologies and classification systems**
  - Published subjects, DAML+OIL, LCSH, DDC
  - Sources of topic types and association types for various domains
- **Database schemas**
  - Entities map to topic types
  - Relations map to association types
- **DTDs and XML schemas**
  - Some element and attribute types map directly to topic types
  - Association types may be inferred from content models
- **Metadata schemas**
  - Mostly provide characteristics for an addressable subject

## Document metadata

- **Properties stored with the file**
  - Word or PowerPoint properties, HTML Dublin Core, PDF-RDF
- **Properties stored externally**
  - RDF, MPEG21, DMS metadata
- **Access is file-format or system specific, but metadata may fall into common categories**
- **Typical metadata values include:**
  - Author, Subject, Keyword
- **Metadata values represent**
  - topics associated with the topic which represents the document, or
  - topics for which this document may be some form of typed occurrence
  - associations between individual values representing topics




## Leveraging existing metadata

```
<item rdf:about="http://www.oreillynet.com/.../metadata.html">
  <title>Distributed Metadata</title>
  <link>http://www.oreillynet.com/.../metadata.html</link>
  <dc:description>This article addresses...</dc:description>
  <dc:subject>metadata, rdf, peer-to-peer</dc:subject>
  <dc:creator>Dan Brickley and Rael Dornfest</dc:creator>
  <dc:publisher>O'Reilly & Associates</dc:publisher>
  <dc:date>2000-10-29T00:34:00+00:00</dc:date>
  <dc:type>article</dc:type>
  <dc:language>en-us</dc:language>
  <dc:format>text/html</dc:format>
  <dc:rights>Copyright 2000, O'Reilly
  & Associates, Inc.</dc:rights>
  ...
</item>
```

*RDF statements about an addressable subject:*

- Topic name and identity
- Potential associations
- Potential occurrences

Dublin Core metadata in RDF format
(from R. Dornfest and D. Brickley, *The Power of Metadata*)

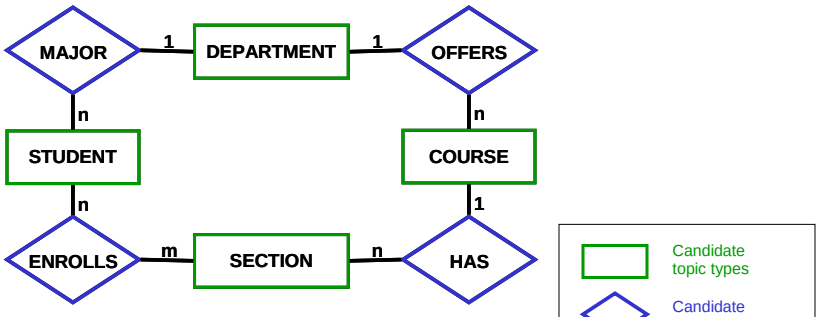
## Information systems

- **Rich repositories of metadata**
- **Access is application or technology specific (RDBMS, LDAP etc)**
- **Topic types and association types are encoded in the schema of the system**
- **A single table often equates to a topic type**
- **Each row in the table is a topic of that type**
- **Columns map to**
  - IDs, names, occurrences or associations with other topics

| ID | Name | Age | Father |
|---|---|---|---|
| 1 | Steve | 48 | 5 |
| 2 | Hedda | 18 | 1 |
| 3 | Lisa | 16 | 1 |
| 4 | Thea | 10 | 1 |




## Leveraging an existing database schema




Entity-relationship diagram for a university database
(from J. Sowa, *Knowledge Representation*)

## Structured document content

- **Semantic markup can be a rich source of topics and associations**
  - The content of certain elements will be the names of topics
  - Associations may be inferred from relations between elements
- **Access is standardised (few file formats)**
  - Enables the same techniques to be applied regardless of document structure
- **May enable a very detailed deconstruction of the content**

```
<address>
   ...
   <city>Oslo</city>
   <country>Norway</country>
</address>
```




## Leveraging structured content

```
<?XML version="1.0"?>
<xmldoc>
<customer>
   <accountid>AE4-Robertson</accountid>
   <name>
      <first>Eric</first>
      <mi>H</mi>
      <last>Robertson</last>
   </name>
   <title>VP Sales</title>
   <contact>
      <wphone>123-555-1212</wphone>
      <hphone>123-555-5678</hphone>
      <email>salesvp@yoyo.com</email>
   </contact>
</customer>
</xmldoc>
```

**Topic types:**
- customer
- account

**Association types:**
- customer-account

**Occurrence types:**
- (name)
- ID?
- title
- work-phone
- home-phone
- email

Data in XML format
(from W.R. Stanek, *Structuring Data with XML*)

## Unstructured document content

- **Extraction depends on analysis of textual content**
- **Natural Language Processing techniques may be applied**
  - Named Entity recognition
    - locating people, companies and places in the text
  - Concept extraction
    - identifying the 'key words' of the text
  - Taxonomic classification
    - analysis according to a human-defined taxonomy
  - Discourse analysis
    - understanding the meaning of the text
- **Access is file-format specific, although most processing tools require only raw text**




## Data extraction techniques

- **A number of approaches are possible:**
  - One time extraction of the topic map
  - Repeated batch extraction process
  - Wrapper around an information system
- **In general, each approach has advantages and disadvantages**
- **In specific cases, one particular approach may have no disadvantages at all**
- **When the legacy data can be subdivided, multiple approaches can be combined through merging**

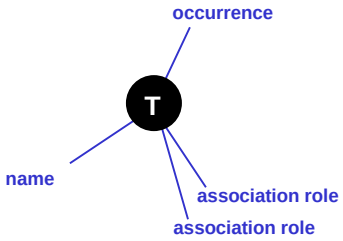
## Merging topic maps

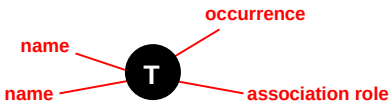
- **Merging has been a central design goal right from the inception of the topic map paradigm**
  - The original motivation was the desire to be able to merge back-of-book indexes
- **A topic map application may use multiple topic maps**
  - Each topic map may emanate from a different source
  - Each topic map may be generated by a different technique
  - Each topic map may even be in a different syntax
- **Merging takes places on the basis of**
  - subject identity (the more robust mechanism)
  - topic names (a less robust fallback)

## Principles of merging (1)

- **When two topic maps are merged, topics that represent the *same subject* should be merged to a *single topic***



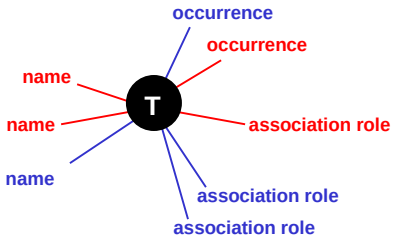




A second topic "about" the same subject

## Principles of merging (2)

- **When two topics are merged, the resulting topic has the *union of the characteristics* of the two original topics**







## One-off extraction (migration)

- **Advantages**
  - Topic map system is legacy-free
  - Full power of topic maps can be used from the start
- **Disadvantages**
  - Risk of 'Big Bang' approach
  - Must either roll-out to all users, or
  - Be prepared to support an increasingly out-dated legacy index as the topic map gets updated.
- **Choose when:**
  - TM project is main business driver
  - TM project is part of a larger project changing the work environment
  - Low risk aversion
  - Legacy data does not lend itself to fully automated conversion

## Batch extraction (scheduled or triggered)

- **Advantages**
  - Continue to use existing tools for creating and updating the source data
  - Topic map structure can be used in parallel with existing (familiar) structures
- **Disadvantages**
  - Requires an explicit conversion phase
  - Conversion routines need updating if schema changes
  - Topic map may not always be up-to-date
- **Choose when:**
  - Significant investment in current toolset
  - Legacy data contains metadata sufficient for automatic topic map generation
  - In an update/review/publish cycle
  - Access to topic map is not critical for creators of information

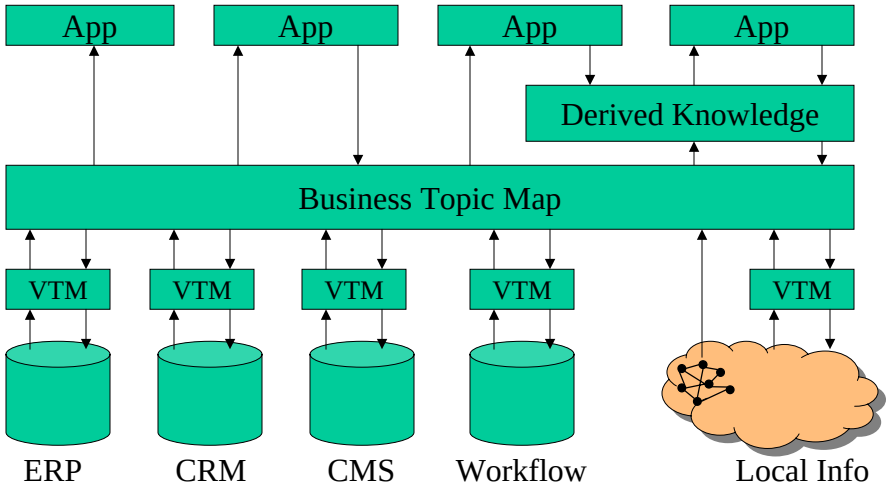
 

## Information system wrapper

- **Advantages**
  - Can continue to use familiar tools and navigational structures
  - Topic map is always up-to-date
- **Disadvantages**
  - More complex than conversion script
  - Significant application changes may be required if the information system schema changes
- **Choose when:**
  - Significant investment in current toolset
  - Legacy systems which do not lend themselves to scripted conversion:
    - Databases (RDBMS, LDAP etc)
    - Document Management Systems
    - Financial Systems
  - Asynchronous / real-time updates required
  - Access to topic map is critical to creators of information

## Virtual Topic Map architecture




## Connecting to occurrences

- **Occurrences may be:**
  - Links back into the data which gave rise to the topics
    - as easy (or difficult) as the task of discovering the topics themselves
  - Links into new data containing mentions of already existing topics
    - considerably easier, since knowledge already in the topic map can be leveraged (e.g. names and types of topics)

## Conclusion

- **Topic map creation is not as difficult as you think**
- **There are many sources of topic map data**
- **Much of this may be leveraged very efficiently through automated processes**
- **Some of it can be mapped directly to topic maps**
- **Manual enrichment adds considerable value**
- **Most topic map creation will be a combination of autogeneration and manual enrichment**




## Want to know more about topic maps?

- **Ontopia Web Site: http://www.ontopia.net**
  - Literature and references
  - Online demos
- **Download Ontopia's free Omnigator**
  - Play with the enclosed topic maps
  - Experiment with your own topic maps (step-by-step tutorial included!)
- **Call the Ontopians for consultancy and training**
  - +47 23233080
  - info@ontopia.net