Automated Metadata Generation and the Critical Role of Catalogers and Indexers in Technical Services of the Future

Denise A. D. Bedford, Ph.d.
Goodyear Professor of Knowledge Management
Kent State University
Kent Ohio

Technical Services Renaissance, Ohio Valley Technical Services Group
Miami University, May 16, 2011
Five Questions

- Making the Case: Catalogers as Knowledge Engineers
- The Future is Semantic
- Automated Metadata Generation – What is it and How does it Work?
- The Role of Human Knowledge in Automated Metadata Generation
CATALOGERS AS KNOWLEDGE ENGINEERS
Little History

- 1979 Director’s Office at Stanford University Libraries

- Cataloging Backlog Analysis and Resolution

- Workflow Investigations, Technical Services Organizational Structure Review ……

- Only way to reduce the backlog and prevent its continual growth was to reduce the unit processing time for cataloging
Little More History …..

- University of California Berkeley Ph.d. program and Stanford courses in operations research, artificial intelligence and natural language processing, and programming-engineering systems

- U. C. Systemwide Administration saw the same continuous growth in number of resources requiring metadata – first online catalogs

- Advances in semantic analysis methods (early 1980’s through 2000’s) and natural language processing

- Increased value of and demand for metadata to support information management and access due to the emerging semantic web
Early Fascination with Natural Language Processing

- In the 1980s, I was part of the community that looked to natural language processing to produce significant improvements in all aspects of information management and access.

- It soon became clear to me, though, that most of the technologies were not going to get us where we needed to be.

- Most of the technologies either used a statistical approach or took a simplistic approach to leveraging Knowledge Organization Systems.

- Different approach was called for ..... What was that approach?
Today’s Dynamic Information Landscape

- Demand for metadata is increasing exponentially today
  - for richer and more granular metadata
  - more resources and more types of resources to process

- Personalization is also a growing factor

- Focus on “content” not just the “package” – which brings design opportunities

- Knowledge is not static – neither is any reference source or knowledge base – need to continuously update our sources
Meeting the Challenge

Automated Metadata Generation allows us to:

- Increase number of resources that have metadata
- Increase the availability of metadata – at the whole and also at the part level
- Increase the number of values for metadata attributes
- Increase the number of attributes (i.e., access points)
- Decrease the time devoted to creating metadata (avg. time from 20 mins. to 2 seconds)
- Improve the quality and consistency of metadata generated
- Meet the increasing demand for personalized views of information
But, It Doesn’t Happen “Automagically”

- Each productive use of technology requires use of existing human knowledge - there is no such thing as a technology that works well “automagically” without human training or design.

- And, no single technology is suited to any or all knowledge processing challenges – each knowledge processing challenge requires that we stop and think about how humans do the task – in order to model how the technology will support it.

- Beware of the “I have a hammer, so everything looks like a nail” syndrome.

- There are also different levels of support – some solutions may be fully automated, whereas others may simply provide assistance to the person performing the task.
The Cost – Teaching Technology to be Smart

- Technology can only behave intelligently – by human standards – when it has human intelligence to work with. Just because technology produces a result doesn’t mean that it is a good result.

- Challenge we face in making technology smart is figuring out (1) how to teach technology what we know and (2) how we think about things.

- Artificial intelligence, psychology, philosophy, communications, education – all have contributed to our understanding of what technology is and is not capable of doing.

- People share what they know, express what they know and record what they know using language – to process information, we need to start at the point of teaching technology how to understand language.
Part 2

THE FUTURE IS SEMANTIC
Semantic Analysis

- Semantic only means that there is some “meaningful” and “understandable” approach involved to solving the problem – can be performed by people and machines

- Relies on formal models or representations of knowledge of language and leverages knowledge of phonology, phonetics, morphology, syntax, semantics, pragmatics and discourse

- Formal models used to capture knowledge include state machines, formal rule systems, logic and probabilistic models

- The foundations of technology based semantic analysis lie in computer science, linguistics, mathematics, electrical engineering and psychology
It’s All About Semantic Analysis

• Good automated metadata generation is grounded in quality semantic analysis

• Semantic analysis can be performed by both people and machines. In both cases, the problem being analyzed is by definition defined by a human expert.

• Always model the human process
  • People have a rich store of linguistic and domain knowledge to draw upon
  • Computers need to be able to have all of that linguistic and domain knowledge encoded and also the rules for using that knowledge
Step 1: Natural Language Processing

This is what a computer does to get to the level of understanding where it can take and act upon our instructions

“The process of assigning a part-of-speech or other lexical class marker to each word in a corpus” [or text] (Jurafsky and Martin)
Part of Speech Tagging

- In order to POS tag content, we need to have a framework or set of tags.

- The tagset should include all possible combinations of category values for a given language. A tagset is generally represented by a string of letters or digits:
  - NNS (gen. noun, plural)
  - AAMP3----2A---- (gen. Adj., Masc., Pl., 3rd case (dative), comparative (2nd degree of comparison), Affirmative (no negation))

- Sample tagsets include those developed at Brown, Penn, Multext.
## Xerox Tagset

<table>
<thead>
<tr>
<th>WORD</th>
<th>LEMMA</th>
<th>TAG</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>the</td>
<td>+DET</td>
</tr>
<tr>
<td>girl</td>
<td>girl</td>
<td>+NOUN</td>
</tr>
<tr>
<td>kissed</td>
<td>kiss</td>
<td>+VPAST</td>
</tr>
<tr>
<td>the</td>
<td>the</td>
<td>+DET</td>
</tr>
<tr>
<td>boy</td>
<td>boy</td>
<td>+NOUN</td>
</tr>
<tr>
<td>on</td>
<td>on</td>
<td>+PREP</td>
</tr>
<tr>
<td>the</td>
<td>the</td>
<td>+DET</td>
</tr>
<tr>
<td>cheek</td>
<td>cheek</td>
<td>+NOUN</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word</th>
<th>POS</th>
<th>Additional POS features</th>
</tr>
</thead>
<tbody>
<tr>
<td>smaller</td>
<td>ADJ</td>
<td>COMPARATIVE</td>
</tr>
<tr>
<td>entire</td>
<td>ADJ</td>
<td>ABSOLUTE ATTRIBUTIVE</td>
</tr>
<tr>
<td>fast</td>
<td>ADV</td>
<td>SUPERLATIVE</td>
</tr>
<tr>
<td>that</td>
<td>DET</td>
<td>CENTRAL DEMONSTRATIVE SG</td>
</tr>
<tr>
<td>all</td>
<td>DET</td>
<td>PREDETERMINER SG/PL QUANTIFIER</td>
</tr>
<tr>
<td>dog’s</td>
<td>N</td>
<td>GENITIVE SG</td>
</tr>
<tr>
<td>furniture</td>
<td>N</td>
<td>NOMINATIVE SG NOINDEFDETERMINER</td>
</tr>
<tr>
<td>one-third</td>
<td>NUM</td>
<td>SG</td>
</tr>
<tr>
<td>she</td>
<td>PRON</td>
<td>PERSONAL FEMININE NOMINATIVE SG3</td>
</tr>
<tr>
<td>show</td>
<td>V</td>
<td>IMPERATIVE VFIN</td>
</tr>
<tr>
<td>show</td>
<td>V</td>
<td>PRESENT -SG3 VFIN</td>
</tr>
<tr>
<td>show</td>
<td>N</td>
<td>NOMINATIVE SG</td>
</tr>
<tr>
<td>shown</td>
<td>PCP2</td>
<td>SVOO SVO SV</td>
</tr>
<tr>
<td>occurred</td>
<td>PCP2</td>
<td>SV</td>
</tr>
<tr>
<td>occurred</td>
<td>V</td>
<td>PAST VFIN SV</td>
</tr>
</tbody>
</table>
POS Tagging Example

```
canon_dict
ref+0001/00/00    DATE
William Clinton   PERSON
lame duck         UTERM

next year

'm' 'm' 'm' 's' 'm'
After next year

'm'
President Clinton

'm'
lame duck

morph_dict
after    P
next     A
year     N
president N
Clinton
be       V
a        A
lame     A
duck     N,V

years

presidents

is are was were

ducks ducked ducking
```
Step 2: Building the Knowledge Base(s)

- Catalogers use many different sources of knowledge to make decisions, to reason about issues, to determine what next step to take in the process, and even when to discard knowledge.

- A cataloger’s underlying tacit knowledge must be integrated into a system that generates metadata automatically if the process is to be performed as effectively by technology as by a person.

- The design challenge here is a significant one – simply representing a word, or a concept or linking concepts in a structure does not assume it can be effectively used by a computer – neither is simply plugging in a thesaurus or classification scheme the same as a “cataloger’s brain.”
How People Classify

- Let’s go back to the most important question – how does a cataloger do it?

- First, we develop knowledge of the classification scheme to which we’re classifying - the better a person’s knowledge of the scheme and the better their knowledge of the object, the better judgment they can make.

- Second, we analyze the object that we’re classifying.

- Third, we make a judgment as to the best fit of the object we’re classifying to all the classes that are available to us –
Caution About Some Technologies

- Rule based classification implies that we have a scheme and defined classes to which to assign entities or objects

- This is a different process than defining classes to constitute a classification scheme – most of the tools do this today

- Much of the “semantic analysis” literature focuses on how to define classes from a set of information – deductively – and then to classify the entities in that set back to the scheme
How a Machine Selects a Class

- From the choices we give them, based on what we tell them about the choices, and the rules we give them to make the selection

- They will choose poorly,
  - if we give them a poorly defined or unbalanced scheme
  - if we tell them nothing or very little about the classes
  - If the manual rules are not rigorous

- You may be surprised to find how often a cataloger is subconsciously compensating for a poorly formed classification scheme…..
A Real Life Example: 
Topic Classification Scheme

Browse - By Topic

- Agriculture
- Communities and Human Settlements
- Conflict and Development
- Culture and Development
- Education
- Energy
- Environment
- Finance and Financial Sector Development
- Gender
- Governance
- Health, Nutrition and Population
- Industry
- Informatics
- Information and Communication Technologies
- Infrastructure Economics and Finance
- International Economics and Trade
- Law and Development
- Macroeconomics and Economic Growth
- Poverty Reduction
- Private Sector Development
- Public Sector Development
- Rural Development
- Science and Technology Development
- Social Development
- Social Protections and Labor
- Transport
- Urban Development
- Water Resources
- Water Supply and Sanitation

- Adaptation to Climate Change
- Air Quality & Clean Air
- Biodiversity
- Brown Issues and Health
- Carbon Policy and Trading
- Climate Change and Environment
- Climate Change Impacts
- Climate Change Mitigation and Green House Gases
- Coastal and Marine Environment
- Drylands & Desertification
- Ecosystems and Natural Habitats
- Environment and Energy Efficiency
- Environmental Disasters & Degradation
- Environmental Economics & Policies
- Environmental Engineering
- Environmental Governance
- Environmental Information Systems
- Environmental Management
- Environmental Protection
- Environmentally Protected Areas
- Forests and Forestry
- Global Environment Facility
- Green Issues
- Marine Environment
- Montreal Protocol
- Natural Disasters
- Natural Resources Management
- Persistent Organic Pollutants
- Pollution Management & Control
- Sustainable Land Management
- Tourism and Ecotourism
- Water Resources Management
- Wildlife Resources
Topic Hierarchy From Relationships across data classes

Build the rules at the lowest level of categorization
Sample Definition of Subclass
Climate Change and Environment

Sample Definition of Subclass
Livestock and Animal Husbandry
Sample Definition of Subclass
Primary Education
Sample Sentiment Analysis Profile

- Negative Sentiment
  - "abandoned"
  - "abashed"
  - "aberrant"
  - "abhorrent"
  - "abject"
  - "abjure"
  - "abolish"
  - "abolishing"
  - "abortive"
  - "absurd"

- Positive Sentiment
  - "admirably"
  - "adorably"
  - "adoringly"
  - "adroitly"
  - "adulately"
  - "advantageously"
  - "adventurously"
  - "aesthetically"
  - "affably"
Another Example: Country Categorization and City Extraction
Operator and Condition Based Matching

Common Matching Operators

AND
OR
NOT
MIN_
DIST_
MINOC_
MAXOC_
START_
END_
ORD
SENT
PAR
NOTIN
NOTINSENT
NOTINPAR
ORDDIST_
MAXPAR_
MAXSENT_
PARPOS_
NOTINDIST_

DIST_200
"World Bank"
"abundantly"

DIST_200
"World Bank"
"acceptingly"

DIST_200
"World Bank"
"accessibly"

DIST_200
"World Bank"
"acclamatorily"

MINOC_2
"Bank accounting"
MINOC_2
"Bank accounts"
OR
"Bank acquisitions"
OR
"Bank acquisitions & me"
MINOC_10
"Bank activity"
MINOC_10
"Bank assets"
OR
"Bank assistance to police"
OR
"Bank automation"
OR
"Bank bailouts"
MINOC_7
"Bank bonds"
OR
"Bank branch offices"
OR
"Bank branches"
MINOC_10
"Bank capital"

Do not match on this concept unless there are a minimum of 10 occurrences in the entity.

If you find this word within 200 characters of “World Bank” then score as one match
Example 3:
Partial Grammatical Concept Extraction for Titles

Full profile is about 4 pages long
Example 4:
ISBN Concept Extraction Profile
Example 5:
People Profile With Authority File of First Names

```
# ROOT=*Name

# This profile is modeled on the new Person Name profile, basi
*Name = *FN1 #cap
*Name = *FN1 #cap #cap
*Name = *FN1 #cap - #cap
*Name = *FN1 *FN1 #cap
*Name = *FN1 __MIDDLEINITIAL #cap
*Name = __MIDDLEINITIAL __MIDDLEINITIAL __MIDDLEINITIAL #cap
*Name = __MIDDLEINITIAL __MIDDLEINITIAL __MIDDLEINITIAL #cap
*Name = *FN1 De #cap
*Name = *FN1 de #cap
*Name = *FN1 da #cap
*Name = *FN1 Da #cap
*Name = *FN1 de la #cap
*Name = *FN1 De la #cap
*Name = *FN1 Del Mar #cap
*Name = *FN1 du #cap
*Name = *FN1 du #cap
*Name = *FN1 du #cap
*Name = *FN1 von #cap
*Name = *FN1 ibn #cap
*Name = *FN1 ben #cap
*Name = *FN1 von #cap
*Name = *FN1 dc #cap
*Name = *FN1 van #cap
*Name = *FN1 van dc #cap
*Name = *FN1 van der #cap
*Name = *FN1 al #cap
*Name = Mr. #cap
*Name = Mrs. #cap
*Name = Ms. #cap
*Name = Miss #cap
*Name = M. #cap
*Name = Mme. #cap
*Name = Me. #cap
*Name = Mr #cap
*Name = Mrs #cap
*Name = Ms #cap
*Name = Mme #cap
*Name = Me #cap
```

# Be certain to include names

*FN1 = *FN
*FN = 'Å,'Kabsila
*FN = 'Aadam
*FN = 'Aadarshini
*FN = 'Aadel
*FN = 'Aadi
*FN = 'Aadil
*FN = 'Aadilah
*FN = 'Aaditya
*FN = 'AE' 'amonn
*FN = 'Aafke
*FN = 'Anfreed
*FN = 'Age
*FN = 'Aghaa
*FN = 'Aakanksha
*FN = 'Aakarshan
*FN = 'Aakif
*FN = 'Aalam
*FN = 'Aaleyah
*FN = 'Aalif
*FN = 'Aali
*FN = 'Aaliyeh
*FN = 'Aamaal
*FN = 'Aaman
*FN = 'Aamil
*FN = 'Aamina
*FN = 'Amir
*FN = 'Anchal
*FN = 'Aagaa
*FN = 'Aara
*FN = 'Aaralyn
*FN = 'Aarif
*FN = 'Aariz
*FN = 'Aaroon
*FN = 'Aarre
*FN = 'Aart
*FN = 'Aarthy
*FN = 'Aarti
*FN = 'Aaryn
*FN = 'Aasaf
*FN = 'Aashish
*FN = 'Aoshiyona
*FN = 'Aashka
```
Classifier concept extraction allows us to look for exact string matches. This requires an exhaustive list—but gives us extensive control. (It would be difficult to distinguish by pattern between IGOs and other NGOs.)
Another list of entities matches exact strings. In this case, though, we’re making this into an ‘authority control list’—we’re matching multiple strings to the one approved output. (In this case, the AACR2-approved edition statement.)
THE ROLE OF HUMAN KNOWLEDGE IN METADATA GENERATION
No Semantic Future Without Catalogers.....

- Catalogers need to be involved in configuring and designing the semantic applications
  - Identifying the best sources of reference knowledge
  - Serving as the “experts” for “expert systems” development
  - Performing quality control on processes

- In the future, catalogers’ knowledge and ways of thinking and working will be the basis of well designed semantic analysis applications

- Both the need for catalogers and the role they play will become critical in the future
Catalogers as Knowledge Engineers

- Role of the cataloger will be shifted in the future from a “Doer” to a “Designer” -- “Knowledge Engineer”

- Designing the context, the content and taking a more proactive role in engineering access to not only information but knowledge

- Future information landscape is inherently “semantic” which aligns very closely with a cataloger’s tacit knowledge

- Cataloger’s tacit knowledge includes rules of thumb, interpretation of guidelines, knowledge of sources, and knowledge of domains
Catalogers as Knowledge Engineers

This shift will mean:

- Learning how to design and build the reference sources, how to develop and apply guiding principles and how to manage reference sources
- Teaching semantic analysis methods and knowledge organization systems
- Putting the tools in the hands of catalogers
- Involving catalogers in the semantic analysis design and development process

In many professional schools, we only teach catalogers how to “use” general purpose reference sources – that source is designed for one area of practice and one general audience - this does not fully leverage our professional knowledge
THANK YOU!

QUESTIONS & COMMENTS?

dbedfor3@kent.edu
dbedford@worldbank.org