Metadata Issues

Sharon Flank
eMotion, Inc.
Overview

- The value of metadata
- Possible structures for metadata
- Dynamic Semantic Expansion
- Managing Hierarchical Information
- Optimizing for the long term
- What’s wrong with keywords?
Why Metadata?

• Image recognition is immature
  – Colors
  – Fingerprint matching

• Transcripts do not speak for themselves
  – Creation info
  – History
  – Implicit info
Possible Metadata Structures

• Fielded Data
  – Automatically collected
  – Added by a human indexer
  – Checkboxes, fill-ins, pull-downs

• Descriptive Data

• Channel Data
  – Closed caption
  – Emerging types (e.g. KLV)
Fielded Data

- Attribute-value pairs
- Independent of database and data structure
- Various search methods
  - Exact match -- Wildcard
  - Keyword -- Numeric
  - Boolean -- Date
- May have “smart” fields
Smart Features

- Synonym matching
  - puma, mountain lion, cougar
- Related term matching
  - big cat, carnivore
- Name identification
  - Hilary, Hillary, H.R. Clinton, George W. Bush, Al, Albert, Gore not gore, Bush not bush
More Smart Features

• Word roots/stemming
  – geese/goose, not geeses or gooses
• Combines with fielded search
  – Can specify categories with a pull-down choice
  – Can specify dates, locations, equipment
  – Can include free-form query text at the same time
More Smart Features

• Locations

• Relative frequency
  – Greater weight on rarer terms

• Phrase recognition
  – fire engine vs. engine fire
  – tiger should not retrieve tiger shark
### Smart Features: Summary

<table>
<thead>
<tr>
<th>Metadata</th>
<th>Techniques Required to Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>daffodil blooming in the snow</td>
<td>semantic net linking <em>flower</em> to subtype <em>daffodil</em></td>
</tr>
<tr>
<td>flower blooms in the snow</td>
<td>morphological analyzer linking <em>blooming</em> to <em>blossoms</em></td>
</tr>
<tr>
<td>flower blossoming in the snow</td>
<td>semantic net linking <em>blooming</em> to synonym <em>blossoming</em></td>
</tr>
<tr>
<td>flower girl dressed as Snow White</td>
<td>Do Not Match: noun phrase analyzer shows <em>flower</em> and <em>snow</em> as modifiers</td>
</tr>
<tr>
<td>flower dusted with snowflakes</td>
<td>semantic net linking <em>snow</em> to component <em>snowflakes</em></td>
</tr>
</tbody>
</table>
Dynamic Semantic Expansion

- Places burden on system, not on cataloguer or searcher
- Based on standard resources
- 400,000+ terms
- Extensible
  - New terms
  - New relationships
  - Multilingual
Managing Hierarchical Information

- Fields vs. objects
- Object model
  - Can represent more complex information
  - Brittle
  - Difficult to maintain, extend
- Fields
  - Simpler
  - Can represent most information efficiently
Objects

- Person
  - Talent
    - Actress: Barbra Streisand
  - Composer
    - Singer: Barbra Streisand
  - Director
    - Barbra Streisand
  - Producer
    - Barbra Streisand
Object Model: Complications

For every new role Barbra Streisand plays:
• Thesaurus manager enters each new role in thesaurus.
• Cataloguers then enter Barbra Streisand (in new role) in metadata.
• No variants permitted.
• Search processing requires traversing the hierarchy, i.e. multiple actions.
Fields

**Actor/Actress:** Barbra Streisand  
**Singer:** Barbra Streisand  
**Producer:** Barbra Streisand  
**Director:** Barbra Streisand  
**Composer:** Barbra Streisand
Field Model

• Cataloguers enter name.
• Variants automatically handled.
  – Barbara, Barb, Ms. B. Streisand
• Search processing is one simple transaction.
Optimizing for the Long Term

- Select a standard that is
  - Long-lasting
  - Well known
  - Easy to learn
## Comparing Standards

<table>
<thead>
<tr>
<th></th>
<th>Home-grown</th>
<th>Library of Congress</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long-lasting</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Well known</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Easy to learn</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>
Retrieval of Keywords

- **Sparse**
  - <5 keywords
  - Boolean keyword search is fine
  - Many assets may have the same keywords
  - May be hard to guess keywords

- **Thorough**
  - >5 keywords
  - Phrases are lost
  - Booleans OK
  - Semantic matching still useful
  - Less need to guess keywords
### Comparing Search Methods

<table>
<thead>
<tr>
<th>Search Method</th>
<th>How Files are Tagged</th>
<th>Easy to Use?</th>
<th>Scales to 1000s of Files?</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural language (plain English)</td>
<td>Plain English caption (2-3 sentences)</td>
<td>Yes</td>
<td>Yes</td>
<td>Over 90% accurate</td>
</tr>
<tr>
<td>Natural language</td>
<td>10 or more keywords</td>
<td>Yes</td>
<td>Yes</td>
<td>Over 80% accurate</td>
</tr>
<tr>
<td>Boolean keyword</td>
<td>10 or more keywords</td>
<td>Somewhat</td>
<td>Yes</td>
<td>50-100% accurate</td>
</tr>
<tr>
<td>Boolean keyword</td>
<td>5 or fewer keywords</td>
<td>Somewhat</td>
<td>No</td>
<td>10-100% accurate</td>
</tr>
<tr>
<td>Natural language</td>
<td>5 or fewer keywords</td>
<td>Yes</td>
<td>No</td>
<td>5-100% accurate</td>
</tr>
<tr>
<td>Fielded search</td>
<td>Fields</td>
<td>No</td>
<td>No</td>
<td>100% accurate</td>
</tr>
<tr>
<td>Classification into topic areas</td>
<td>Topics</td>
<td>No</td>
<td>No</td>
<td>100% accurate</td>
</tr>
</tbody>
</table>
What’s Wrong with Keywords?

• Hard to write
  – Haphazard
  – Need tools to standardize even a little
  – Cataloguers need extensive training, QA

• Hard to maintain
  – Difficult to keep up standards over time

• Hard to search
  – Searchers need to guess at keywords
Why English?

• With Natural Language search, system takes on the burden of relating terms
• 50% cheaper/faster to catalogue
• Easy to train cataloguers
• No search training needed
• Higher precision even for advanced users
XML is Not Enough

• Simply a labeling mechanism
  
  \[<\text{location}>\text{Washington}</\text{location}>\]
  \[<\text{person}>\text{Washington }</\text{person}>\]

  vs.

  \text{Location: Washington}
  
  \text{Person: Washington}

• Doesn’t tell us which one matches George Washington or Washington DC or East
Conclusion

• Select a standard that is
  – Low-maintenance
  – Easy for cataloguers
  – Easy to extend
  – Supportive of naïve users
  – Supportive of expert searchers