Adventures of Categories

Modelling the evolution of categories during scientific investigation

Prashant Gupta, Mark Gahegan and Gill Dobbie
Centre for eResearch
The University of Auckland
New Zealand
An example of iterative and dynamic process of knowledge creation in Geoscience

Change happens in multiple dimensions

- Time
- Space

“Using data from Indonesia, we show that the choice of a forest definition can have a large impact on estimates of deforestation and forest degradation areas, on assessment of drivers of deforestation and on the development of a REL.”


Example – A ‘Forest’ category has almost 800 different definitions

Current state of Knowledge evolution

- Multiple domains working on knowledge evolution
  - Concept drift (machine learning)

What is missing?

- Tools supporting evolution are disconnected from the process of science.

- Missing how and why aspects of change
  - Change logs
  - Provenance
What-ifs

• How we currently deal with the effects of change?
Knowledge producer vs. consumer perspective

Knowledge producer

Knowledge consumer

It's very confusing. Hard to say how they were used. I wish they had some explicit connections.
Iterative, dynamic and exploratory process of evolution

The trajectory [1] of a category not just refers to the category evolution (in terms of static versions), but also embraces the (dynamic) interactions of multiple artifacts, processes and people contributing to the construction and evolution of categories.

Exploration path - the iterative interplay between various methods, data and human understanding before a researcher reaches the final workflow design

AdvoCate – **Adventures of Categories**

- An e-Science tool to support category evolution
- Allows researchers to model changes, captures the exploration paths behind a change, and maintains a category-versioning system
- Support change propagation between multiple systems (database, ontology and maps)
Deeper representation of categories
Conceptual model
Change language

Change Algebra

<table>
<thead>
<tr>
<th>Typology</th>
<th>Operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth</td>
<td>⊗ C</td>
</tr>
<tr>
<td>Death</td>
<td>⊗ C</td>
</tr>
<tr>
<td>Split</td>
<td>C ⊕ C₁, C₂</td>
</tr>
<tr>
<td>Merge</td>
<td>C ⊕ C₁, C₂</td>
</tr>
<tr>
<td>Drift</td>
<td>C ↝ C’</td>
</tr>
</tbody>
</table>

Internal & external change triggers

<table>
<thead>
<tr>
<th>Change trigger (external)</th>
<th>Change trigger (internal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social requirement (addition of a new category)</td>
<td>Change in training sample</td>
</tr>
<tr>
<td>Social requirement (descriptive categories)</td>
<td>Change in training sample</td>
</tr>
<tr>
<td>Scientific requirement (accuracy)</td>
<td>New classifier</td>
</tr>
<tr>
<td>Scientific requirement (accuracy)</td>
<td>Change in training sample</td>
</tr>
<tr>
<td>Scientific requirement (accuracy)</td>
<td>Change in parameters</td>
</tr>
<tr>
<td>Conceptual change</td>
<td>Change in training sample</td>
</tr>
<tr>
<td>Error in data collection activity</td>
<td>Change in training sample</td>
</tr>
</tbody>
</table>
Process model of category evolution

- New observation
- Scientific requirements
- Societal factors
- New understanding

- Change in training samples
- Change in classifier
- Change in parameters

- Elementary changes:
  - Add/Delete category
  - Add/Delete relationship
  - Change label
  - Change intension

- Composite changes:
  - Born
  - Die
  - Merge
  - Split
  - Drift

- New category
- Splitting or merging of categories
- Drift in categories

Implementing the changes & updating category versioning system

Change approval

Change Propagation
Change scenario
**LCDB1**

- **Water**
- **Vegetation**
- **Built Space**
- **Open Space**
- **Cloud**
- **Shadow**

- New training samples with more spectral bands
- Training samples bimodal – added a new category (Forest)
- Updated training samples with Forest category

**LCDB2**

- **Water**
- **Vegetation**
- **Indigenous Forest**
- **Urban**
- **Suburban**

- New training samples with better spatial resolution
- New bands help to identify Urban and Suburban areas. We split Built Space into sub-categories

**LCDB3**

- **Salt Water**
- **Fresh Water**
- **Vegetation**
- **Indigenous Forest**
- **Urban**
- **Suburban**

- Split Water category into Salt Water and Fresh Water
- Used new SVM classifier – better accuracy

**Activities**

- A1 New training samples collection
- A2 Edit training samples
- A3 Create signature file
- A4 Classification

**Symbols**

- Comment related to an activity
- Control flow
- Path resulting to the final category
Conclusion

- Supports richer representation of categories and the evolution process
- Bridges the gap between the process and products of science – highlighting the dynamic aspect of scientific knowledge
- Connects and synchronizes various knowledge components through the process of change
Questions?

@pguptanz

p.gupta@auckland.ac.nz
Deeper representation of categories

Deeper representation of categories

**Concept (mental)**

**Category**

**Intension**

**Categorical model**

**Extension**

**Training data**

**Exploration paths**

**Relationships**

- **e.g. Forest** – a land cover category

- **Probable distribution model** - identifies whether a pixel belongs to 'Forest'

- **Prototype pixels** – exemplify the 'Forest' category in a remote sensing image

- **Child of**: Agricultural Land
  **Parent of**: Exotic Pine
Change language

Change Algebra

<table>
<thead>
<tr>
<th>Typology</th>
<th>Operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth</td>
<td>⊗ C</td>
</tr>
<tr>
<td>Death</td>
<td>⊗ C</td>
</tr>
<tr>
<td>Split</td>
<td>C ⊕ C₁, C₂</td>
</tr>
<tr>
<td>Merge</td>
<td>C ⊕ C₁, C₂</td>
</tr>
<tr>
<td>Drift</td>
<td>C ↳ C'</td>
</tr>
</tbody>
</table>

Internal & external change triggers

<table>
<thead>
<tr>
<th>Change trigger (external)</th>
<th>Change trigger (internal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social requirement (addition of a new category)</td>
<td>Change in training sample</td>
</tr>
<tr>
<td>Social requirement (descriptive categories)</td>
<td>Change in training sample</td>
</tr>
<tr>
<td>Scientific requirement (accuracy)</td>
<td>New classifier</td>
</tr>
<tr>
<td>Scientific requirement (accuracy)</td>
<td>Change in training sample</td>
</tr>
<tr>
<td>Scientific requirement (accuracy)</td>
<td>Change in parameters</td>
</tr>
<tr>
<td>Conceptual change</td>
<td>Change in training sample</td>
</tr>
<tr>
<td>Error in data collection activity</td>
<td>Change in training sample</td>
</tr>
</tbody>
</table>

Relational Operations

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is same as</td>
<td>≍</td>
</tr>
<tr>
<td>contains</td>
<td>⊂</td>
</tr>
<tr>
<td>Is contained by</td>
<td>⊆</td>
</tr>
<tr>
<td>Is confused with</td>
<td>∥</td>
</tr>
<tr>
<td>Is independent of</td>
<td>⊬</td>
</tr>
</tbody>
</table>
How and why were more specialized categories of Water – Salt Water and Fresh Water – modelled in 2008 and not in 2004?
Process model of category evolution

- New observation
- Scientific requirements
- Societal factors
- New understanding

- Change in training samples
- Change in classifier
- Change in parameters

- Elementary changes:
  - Add/Delete category
  - Add/Delete relationship
  - Change label
  - Change intension

- Composite changes:
  - Born
  - Die
  - Merge
  - Split
  - Drift

- New category
- Splitting or merging of categories
- Drift in categories

Process of science

- Change in training samples
- Change in classifier
- Change in parameters

Implementing the changes & updating category versioning system

Change Propagation

Ontology/database evolution tools

Change approval