Categories are in flux, but their computational representations are static and isolated: that’s a problem

Mark Gahegan & Prashant Gupta
Computer Science and Centre for eResearch
The University of Auckland
Ontology: definition

Asking the questions only a child would ask…

…And answering them in the language only a lawyer would use!
“According to Heraclitus, *panta rhei* — everything is in flux. But what gives that flux its form is the *logos* — the words or signs that enable us to perceive patterns in the flux, remember them, talk about them, and take action upon them even while we ourselves are part of the flux we are acting in and on...

Any system of ontology that is adequate for defining the concepts used in natural languages must be at least as flexible as the languages themselves”  
John Sowa: Signs, Processes and Language Games
Categories

• Like Heraclitus, I think of objects and their categories as simply the instantaneous snapshot of the outworking of processes.

• These processes occur at different levels of abstraction, for example:
  • there are natural processes (and societal processes) occurring in our world
  • We see them through observation processes
  • We interpret them through analytical processes
  • We gain understanding of them through experiential processes
  • We agree how and what we will name and describe via social processes
Die Data Science processes Contexts/ Situations Researcher’s experience

Birth of a category

Theory

Category
Intension
Extension
Position in conceptual hierarchy

Conceptual change
New observations
Societal concerns
New computing techniques

Evolution

Die

Category
Intension
Extension
Position in conceptual hierarchy
Fragmentation of scientific artifacts and processes among communities

Community 1: Remote sensing system
- Satellite Imagery
- Data observation
- New/better technology

Community 2: Image processing tools
- Image processing
- Digitized data
- Richer/better data

Community 3: New validation techniques
- Aerial photography
- Field work
- CSV/XML/database
- Concepts
- New ideas

Community 4: Classification
- Machine learning tool
- Land-cover dataset
- Conceptual change

Community 5: Web-mapping tool
- Land-cover map
- Taxonomy
- Applications
- Taxonomy tool
- New/better technology
Interlinking methods, models, data, samples..

[Diagram showing interlinking between data, metadata, models, external databases, and articles with associated URLs: http://www.seek4science.org and http://www.isatools.org]
Other knowledge integration models..

• **Research Objects**

• **Reproducible Research System**

• **Linked Science**

• **Workflows**

  What are the shortcomings?

  • Focus on a single experiment of science, rather than science as an ongoing and evolving process
  • Provide a linear view of science, but science is instead exploratory, dynamic and cyclic
  • Focus typically on data and not on conceptual structures
Connecting scientific artifacts

Conceptual model of knowledge evolution

- Essentially a model of the science process
- Live connections among scientific artifacts

Categories

Database

Ontology

Data

Map

Software tools
Adventures of Categories (AdvoCate)

• An integrated system for managing categories in action, based on a process model of category evolution
• Captures changes in categories, via the process of category evolution and maintains a category-versioning system
• Allows the entire evolution process to be replayed, questioned, communicated
• Can compare versions of categories based on intension as well as extension
Process model of category evolution

- New observation (training data)
- Societal drivers
- New understanding

Implementing the changes & updating category versioning system

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

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

Revising categorical model

- Evaluation of categorical model

Change report (using elementary and complex change operations)

Change approval

Process of science

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

• New observation
• Societal drivers
• New understanding

Ontology/database evolution tools

Change Propagation
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>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>
An example of category evolution from land cover mapping
LCDB1

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

New training samples with more spectral bands

A1 → A2 → A3 → A4

A2

Split Water category into Salt Water and Fresh Water

A1

High error - no change conceptualized

Tried new classifier

A1

A3

A4

A4

Used new SVM classifier – better accuracy

LCDB2

- Water
- Vegetation
- Indigenou sForest
- Urban
- Suburban
- Open Space
- Cloud
- Shadow

New training samples with better spatial resolution

A1 → A2 → A3 → A4

A1

A3

A4

A4

Used new SVM classifier – better accuracy

LCDB3

- Salt Water
- Fresh Water
- Indigenou sForest
- Urban
- Suburban
- Open Space
- Cloud
- Shadow

New training samples with better spatial resolution

A1 → A2 → A3 → A4

A1

A3

A4

A4

path resulting to the final category

Activities

- New training samples collection
- Edit training samples
- Create signature file
- Classification

Year

2000
2004
2008

Symbols

Comment related to an activity

Control flow

path resulting to the final category
Note: The taxonomy is modeled using a Naive Bayes classification model with an accuracy of 90.0%.
The current exploration process, resulting from an external trigger of 'Conceptual change', models changes in existing taxonomy *AKL LCDB* and leads to the following changes.

The new classification model is conceptualized as a Support Vector Machine with an accuracy of 86.0% as compared to the existing classification model stored in AdvoCate as Naive Bayes with an accuracy of 90.0%.

Below is given the different lists of categories and the changes that occurred:

### Comparison between categories corresponding to the common concepts:

<table>
<thead>
<tr>
<th>Concept</th>
<th>User accuracy (new)</th>
<th>Producer accuracy (new)</th>
<th>User accuracy (existing)</th>
<th>Producer accuracy (existing)</th>
<th>Extension similarity</th>
<th>J M Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloud</td>
<td>0.93</td>
<td>1.00</td>
<td>0.95</td>
<td>0.99</td>
<td>0.59</td>
<td>N/A</td>
</tr>
<tr>
<td>Inland Water</td>
<td>0.86</td>
<td>0.88</td>
<td>0.87, 0.82</td>
<td>0.86, 0.84</td>
<td>0.25, 0.36</td>
<td>N/A</td>
</tr>
<tr>
<td>Shadow</td>
<td>0.94</td>
<td>0.96</td>
<td>0.85, 0.83</td>
<td>0.95, 0.96</td>
<td>0.68, 0.68</td>
<td>N/A</td>
</tr>
<tr>
<td>Suburban</td>
<td>0.94</td>
<td>0.94</td>
<td>0.85</td>
<td>0.9</td>
<td>0.72</td>
<td>N/A</td>
</tr>
<tr>
<td>Urban</td>
<td>0.96</td>
<td>0.92</td>
<td>0.86</td>
<td>0.83</td>
<td>0.48</td>
<td>N/A</td>
</tr>
</tbody>
</table>

### Details of the categories corresponding to the concepts resulted from splitting an existing concept:

<table>
<thead>
<tr>
<th>New concept</th>
<th>User accuracy</th>
<th>Producer accuracy</th>
<th>Split from</th>
<th>Extension containment</th>
<th>J-M distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grassland</td>
<td>0.74</td>
<td>0.80</td>
<td>Pasture</td>
<td>0.66</td>
<td>N/A</td>
</tr>
<tr>
<td>Open Space</td>
<td>0.62</td>
<td>0.76</td>
<td>Pasture</td>
<td>0.52</td>
<td>N/A</td>
</tr>
<tr>
<td>Sea Water</td>
<td>0.87</td>
<td>0.92</td>
<td>Water</td>
<td>0.97, 0.99</td>
<td>N/A</td>
</tr>
<tr>
<td>Estuarine Open Water</td>
<td>0.74</td>
<td>0.60</td>
<td>Water</td>
<td>0.86, 0.84</td>
<td>N/A</td>
</tr>
<tr>
<td>scrub</td>
<td>0.56</td>
<td>0.63</td>
<td>Woody Vegetation</td>
<td>0.58, 0.97</td>
<td>N/A</td>
</tr>
<tr>
<td>Indigenous Forest</td>
<td>0.76</td>
<td>0.75</td>
<td>Woody Vegetation</td>
<td>0.92, 0.92</td>
<td>N/A</td>
</tr>
<tr>
<td>Mangrove</td>
<td>0.98</td>
<td>0.96</td>
<td>Woody Vegetation</td>
<td>0.49, 0.63</td>
<td>N/A</td>
</tr>
</tbody>
</table>

### Grouped concepts:

<table>
<thead>
<tr>
<th>Parent concept</th>
<th>Concepts that are grouped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water Bodies</td>
<td>Estuarine Open Water, Inland Water, Sea Water</td>
</tr>
<tr>
<td>Forest</td>
<td>Indigenous Forest, Mangrove</td>
</tr>
<tr>
<td>Vegetation</td>
<td>Forest, Scrub, Grassland</td>
</tr>
<tr>
<td>Built-up Area</td>
<td>Suburban, Urban</td>
</tr>
<tr>
<td>Artificial Surface</td>
<td>Built-up Area, Open Space</td>
</tr>
</tbody>
</table>

Do you wish to implement these changes in AdvoCate: [Yes] [No]
Logical design

Artifacts

Hierarchical Relationship <T1>

Taxonomy <T1, V>

Classifier

Classification model <T1>

Training set <T1, T2, V>

Classification <T3>

Edit training set <T3>

Create training set <T3>

Clustering <T3>

Activities

Training <T3>

Exploration Path <T1>

Change operations

Create taxonomy

Add Concept

Retire concept

Category instantiation

Merge concepts

Group concepts

Split concepts

Change event <T3>

Concept drift

Structural drift

< T3 >

Versioning
Transaction time period
Valid Time period
Transaction time stamp
Conclusions

• I believe it is possible to engineer systems that contain both the methods for doing science AND a meta-model of the science process, so we can explicitly see how these two worlds connect.

• This bridges the gap between the process and products of science – revealing the dynamic and evolving aspects of knowledge.

• It also connects and synchronizes all of the research artefacts through the process of evolution.
End

Questions, comments?
Sources of ontological confusion
(Gahegan & Brodaric, 2014)

**Conceptual:** geoscientists are using different concepts and categories; mapping of new areas, or scientific evolution, often requires existing concepts to be revised or supplanted in the field.

**Theoretic:** geoscientists may use different theories with the same evidence and categories.

**Inferential:** geoscientists may use different reasoning mechanisms.

**Intentional:** different purposes or goals, including choice of conceptual or geographic scale and level of detail of observation, may naturally lead to diverse results.

**Evidential:** geoscientists are considering different data.

**Model-based:** given the same evidence, concepts, theories, and reasoning techniques open systems such as the Earth’s may still lead to the generation of diverse valid-process models.

**Methodological:** geoscientists may use different methods and instruments or perform different actions leading to diverse models.

**Tacit:** geoscientists’ implicit understanding of the region, developed in concordance with unconscious predispositions, may differ and lead to model variability.

**Social:** knowledge transfer between geoscientists may vary according to the degree of scientific interaction as facilitated or impeded by political, cultural, and institutional or other structures.

**Historical:** geoscientists might still develop diverse models due to the order of presentation of evidence and the sequence of decisions made at each stage of observation and reasoning.
Another MOTIVATING EXAMPLE: Map construction and semantic conflict

C: intra-geologist clustering

C: inter-geologist similarity

Gahegan & Brodaric, 2002, Brodaric & Gahegan 2006
Aside: are we relying too much on ontology as our ‘carrier of meaning’?

*Ontology* tells us what is known, but *epistemology* considers how it is known, how it came to be, and why it came to be the way it is (and not some other way); *pragmatics* addresses how it is understood, who understands it…
Levels of Meaning in systems

- **How?** procedural
- **Why? Why?** Situational
- **What?** Definitional
- **knowledge**
- **wisdom?**
- **What?** information
- **data**
- **Semantic Web**
- **Consensus**
- **Semantics**
- **Schematics**
- **Syntax**
- **Systems**
- **Pragmatics**
- epistemology
- Semantics
- Syntax
- Empirics
- Philosophy
Knowledge: Ontology, Epistemology, Pragmatics

• Classically, **Ontology** describes what we know, or what is true, via description logics

• **Epistemology** describes how we know something is true, via methodology, research paradigms

• **Pragmatics** describes the process of interpretation, how and why humans construct and communicate meaning. It is experiential.