The Environment Ontology - Linking Environmental Data

eEnvironment
Thu 26th March 2009, Praha, Czech Republic

Norman Morrison on behalf of the Environment Ontology Consortium
NERC Environmental Bioinformatics Centre, Oxford
& University of Manchester.
www.environmentontology.org

Welcome to the The Environment Ontology
Overview

- Background
- Environmental data and meta-data
- Introduce some of the work of Environment Ontology (EnvO) Consortium
  - Larger infrastructure
  - Introduction to ‘philosophical’ background to the foundations of EnvO
  - Linking environmental data
  - Early adopters
- Future work
Background

- Environmental data are essential to the prediction and mitigation of:
  - The effects of climate change on biodiversity, agriculture and livestock...
  - The extent and progression of ecological and epidemiological outbreaks
  - The effects of environmental perturbation on human health and disease
  - …

- Global environmental data now represents decades of investment by governmental bodies, commercial businesses and higher education institutions.
Capture of Environmental Data
Capture of environmental data
A wealth of environmental data

- Scientific research is benefiting from an unprecedented explosion of data

  “Environmental data are our greatest asset”

  “Our data holdings over the next year, will double that which has been collected in the history of the British Antarctic Survey”

  - David Walton, Emeritus Fellow, BAS.
At a molecular level

Medium scale metagenomic projects generating Hundreds of Gb of data per year
Data about data

- Key to the federated discovery and retrieval of environmental data is the identification, and capture of meta-data types that represent potential nodes of integration.
Integration via meta-data

- Common to all environmental data is a sample or sampling event from which they were derived.

- **Every** environmental sample or sampling event can be linked to a particular locale.
  - Location can be described by coordinate reference system (GPS)

  It would behove spatial scientists to articulate to the broader research community the potential of recording and making accessible spatial data in the appropriate formats — and the painlessness of the process.


- The environment of this locale can in turn be described as a function of the particular conditions that inhere therein.
Six degrees of separation

To

John Morrison
The Isle of Harris
(I think his brother’s name is Donald)

- 1967 - Psychologist Stanley Milgram Harvard University
  - Asked individuals to send a package to a certain person in Boston
  - Described only by Name, including some general features and the fact that they lived in Boston…
  - 64 of the 300 packages made it to the designated recipients!
Six degrees of separation

- When we specify a query…
  - Think of it like setting up a virtual address
  - The address (or meta-data) on the package is what gets it to it’s destination
  - The difference is, when we annotate a package of environmental information, we don’t always know who the recipient will be
  - In this example, key to the package reaching it’s destination is the extra knowledge associated with the address
  - There has to be some knowledge of how meta-data are inter-related which constrains the logically possible interpretations
- So, if we are looking to receive packets of information based on environmental meta-data information…
  - I would hope that we can do better than 64 out of 300! (21%)
Real world examples of ‘habitat’

- Peregrine Falcon
  - "Peregrines are found worldwide except Antarctica."

- Cultured Microbe
  - “The swimming pool of a Nigerian dictator”.
Common Reference Frameworks

- There is increasing recognition therefore that the scientific community would benefit from the development of a common reference framework for describing environmental information.
- Integration and exchange of data through adequate reporting of metadata requires community-level agreement on the nature and structure of the information to be interchanged.
- This is where the development of controlled vocabularies, thesauri and ontologies can prove pivotal.
- The enormous successes of the Gene Ontology (Ashburner et al. Nature Genetics, 2000) and similar artifacts within the biological domain have shown how powerful semantic retrieval and integration of data can be.
The Environment Ontology Project

- **Aim**
  - To support the semantically consistent description of, and computational reasoning over, environmental information associated with environmental data.

- **How**
  - Through the development of an open-source, community-based Environment Ontology (EnvO); and Gazetteer (Gaz) constructed on ontological principles.

- **Who**
  - Developed by The Environment Ontology Consortium
    - Michael Ashburner - University of Cambridge, UK.
    - Dawn Field - NERC Centre for Ecology and Hydrology, UK.
    - Suzanna Lewis - Lawrence Berkeley Laboratories, US.
    - Lynn Schriml - University of Maryland, US.
    - Barry Smith - University at Buffalo, US.
    - …and many others (see website).
  - Developed as part of the Open Biomedical Ontologies Foundry http://obofoundry.org/
# OBO Foundry (+ Environments)

<table>
<thead>
<tr>
<th>RELATION TO TIME</th>
<th>CONTINUANT</th>
<th>OCCURRENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRANULARITY</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ORGAN AND ORGANISM</th>
<th>INDEPENDENT</th>
<th>DEPENDENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organism (NCBI Taxonomy)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anatomical Entity (FMA, CARO)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CELL AND CELLULAR COMPONENT</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell (CL)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cellular Component (FMA, GO)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MOLECULE</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecule (ChEBI, SO, RnaO, PrO)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
EnvO versus Gaz

- national geopolitical entity
- reef
- city
- river
- bridge

- Azerbaijan
- The Great Barrier Reef
- Paris
- The Mississippi
- Golden Gate Bridge
EnvO versus Gaz

- Approx 1,200 terms
  - Environmental matter
  - Environmental feature
  - Biome
  - Food

- Approx 80,000 terms
  - 1st and 2nd order geographic regions
http://bioportal.bioontology.org/
Example query

What metagenomic data are available from samples taken from freshwater found in East Africa?
Linking Environmental Data

Concept 1  Concept 2  Concept 3

'Freshwater'  'found in'  'East Africa'

- Pond
- Lake
- Puddle
- Stream
- Groundwater
- Tapwater

- Found in
- Located at
- Located in
- Placed in
- From

- Kenya
- Tanzania
- Lake Albert
- Entebbe
- Ethiopia
- Nairobi
- Rwenzori Mountains
Relaxing the search space

<table>
<thead>
<tr>
<th>Concept 1</th>
<th>Concept 2</th>
<th>Concept 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>deep sea thermal vent</td>
<td>Located in</td>
<td>marine habitat</td>
</tr>
<tr>
<td>coral reef atoll</td>
<td>Found in</td>
<td>terrestrial habitat</td>
</tr>
<tr>
<td>oceanic trench</td>
<td>Located at</td>
<td>aerial habitat</td>
</tr>
<tr>
<td>mountain</td>
<td>Located in</td>
<td>arboreal habitat</td>
</tr>
<tr>
<td>air stream</td>
<td>Placed in</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>Located in</td>
<td>...</td>
</tr>
</tbody>
</table>
Early Adopters

- GEMINA gemina.igs.umaryland.edu
- STRAININFO www.straininfo.net
- Encyclopedia Of Life (EOL) www.eol.org
- BOLD www.barcodinglife.org
- MITRE www.mitre.org
- GOLD www.genomesonline.org
- GRAMENE www.gramene.org
- CAMERA camera.calit2.net
- GSC GENOME CATALOGUE gensc.org/gsc/gcat
**Gemina** is a web-based system designed to identify infectious pathogens and their representative genomic sequences through selection of associated epidemiology metadata. Gemina supports the development of DNA signature-based assays for the detection of pathogens or sets of pathogen through the Insignia Signature Pipeline at the University of Maryland. View the Quick Start below or the Gemina Tutorial for help on searching the database.

**Selection Summary**

**Controlled Vocabulary**
- Pathogens
- Toxins
- Hosts
- Reservoirs
- Environmental Sources
- Diseases
- Anatomy
- Transmission Methods
- Symptoms
- Location

**Environmental Sources Tree View**
Search by non-living sources of pathogens. As a member of the EnvO consortium, Gemina is a contributing member to the development of this location ontology shown here. The most up-to-date version of the envo.obo file can be downloaded from the EnvO sourceforge site.

- biome
- environmental feature
  - geographic feature
  - habitat
    - aerial habitat
    - anthropogenic habitat
    - aquatic habitat
    - arboreal habitat
    - extreme habitat
    - organism-associated habitat
    - subterrestrial habitat
    - terrestrial habitat
Conclusions

- Initiatives for digital research infrastructure should focus more on making standardized data openly available, and less on developing new portals (Editorial, Data for the masses, Nature, 2009, 457: 129)

- Making standardized data openly available will spur innovation of superior information services

- Formation of an Open Environmental Ontologies (OEO) Foundry
  - The principals of which will be to promote best practice and coordinated orthogonal development of ontologies within the environmental domain
  - Disseminate best practice in technological and content governance mechanisms for terms
    - Persistent URL’s - PURL
Future Work

- **Software tools**
  - Web accessible free-text ontology mark-up tools

  **Ontology Plugin for Word**
  Microsoft Releases Open Tools to Enhance Scientific Research Efforts
  Building on Science Commons Ontologies
  Breakthrough collaboration helps researchers make easier connections on the Web.

  - EnvO supported in word plugin
Future Work

GAZ Term
South Downs

The Definition of This Term

EnvO Terms Detected in the Definition
chalk
ENVO:00002054
"A soft, white, porous limestone." [MA:ma]

Location Extracted From Wikipedia
Lat/Lng (true): 50.55N, 0.30W
Lat/Lng (decimal): 50.91, -0.50
Future Work
This ontology needs you!

- Only through the annotation of data can we make it work
- Trial and Attribution
  - Give it a trial and we’ll give you attribution
- We are an open access community

www.environmentontology.org
Acknowledgements

- Dawn Field, NERC Centre for Ecology and Hydrology
- Suzanna Lewis, Lawrence Berkeley National Laboratory
- Barry Smith, National Center for Biomedical Ontology / University at Buffalo
- Michael Ashburner, Department of Genetics, University of Cambridge
- Brandon Bennett, University of Leeds
- Tim Booth, NERC Environmental Bioinformatics Centre (NEBC)
- Neil Cathness, Oxford e-Research Centre
- Andrew Cossins, School of Biological Sciences, University of Liverpool
- Peter Dawynt, StrainInfo Bioportal / Ghent University
- Salman Elahi, Freshwater Biological Association
- John Goodwin, Ordnance Survey
- Tanya Gray, NERC Centre for Ecology and Hydrology
- Aaron Gussman, J. Craig Venter Institute
- Neil Hall, School of Biological Sciences, University of Liverpool
- Glen Hart, Ordnance Survey
- Stewart Houten, NERC Environmental Bioinformatics Centre (NEBC)
- Pankaj Jaiswal, Dept. of Plant Breeding and Genetics, Cornell University
- Peter Kille, School of Biosciences, Cardiff University
- Kelvin Li, J. Craig Venter Institute
- Joanne Luciano, MITRE
- Chris Mungall, Lawrence Berkeley National Laboratory
- Neil Sarkar, Marine Biological Laboratory
- Robert Stevens, Ontogenesys Network; School of Computer Science, The University of Manchester, UK
- Matthew Stiff, NERC Centre for Ecology and Hydrology
- Mark Viant, School of Biosciences, University of Birmingham
- Steve Young, Oxford e-Research Centre
- David Shotton, Ontogenesis Network; Image Bioinformatics Research Group, Oxford University
- Jun Zhao, Image Bioinformatics Research Group, Oxford University
- Lynn Schriml, University of Maryland School of Medicine
- Bart van Brabant, StrainInfo Bioportal / Ghent University
- Victor Markowitz, Lawrence Berkeley National Laboratory
- Nikos Kyrpides, DOE JGI / Lawrence Berkeley National Laboratory
- Maureen Donnelly, Center of Excellence in Bioinformatics and Life Sciences / University at Buffalo
- Lynette Hirschman, MITRE
- Lincoln Stein, CSHL
- Stephanie Greene, USDA, ARS National Temperate Forage Legume Genetic Resources Unit
- Sujeevan Ratnasingham, University of Guelph
- Jeff White, USDA-ARS