Mining Biodiversity: Enriching Biodiversity Heritage with Text Mining and Social Media

Sophia Ananiadou
National Centre for Text Mining
School of Computer Science
University of Manchester
The partners
Mining Biodiversity is one of the projects that won in the third round of the transatlantic Digging Into Data Challenge, a competition aiming to promote the development of innovative computational techniques that can be applied to big data in the humanities and social sciences. Overall, 14 teams representing collaborations from Canada, US, UK and the Netherlands have been selected. Four of the selected projects come from the UK.

The project is an international collaboration between the National Centre for Text Mining (UK), Missouri Botanical Garden (US) and Dalhousie University’s Big Data Analytics Institute (Canada) and Social Media Lab (Canada). NaCTeM was also a recipient of the 2011 Digging into Data call with the Integrated Social History Environment for Research (ISHER) project.

Mining Biodiversity starts in March 2014 and is funded by AHRC, ESRC, Innovation.ca, Institute of Museum and Library Services, JISC and NEH until September 2015.
Mining Biodiversity

• Transform BHL into a next-generation social digital library
Biodiversity Heritage Library

- a consortium of botanical and natural history libraries
- stores digitised legacy literature on biodiversity
- currently holds 130,000 volumes = over 40 millions of pages (PDFs and OCR-generated text)
- open-access
Current features

• supports keyword-based search
• species names annotated and linked to the Encyclopaedia of Life
• integrates automatic taxonomic name finding tools (uBio, Taxonfinder)
• data access through export functionalities and Web services
Keyword-based search and Browsing
Document viewing

Biologia Centrali-Americana: zoology, botany and archaeology (edited by Frederick Duc...)

Pages
Page xiv
Page xv
Page xvi
Page xvii
Page xviii
Page xix
Page xx
Page xxi
Page xxii

URL for Current Page
http://biolib.biodiversitylibrary.org/page/529535

Scientific Names on this Page
Acarus
Acanthocentrus
Achates
Aelostomus
Alosa
Asynis
Boleosoma
Cichlasoma
Cichlasoma maculicauda
Cichlasoma altifrons
Geophagus
Notopterus

Viewing Page as Text
This text is generated from uncorrected OCR and as such, may contain, inconsistencies with the actual content of the original page.

Nothing is known of the Cichlid fauna of Honduras, Salvador, and Nicaragua north and west of the Great Lakes, but it may be inferred that it is of the same general character as that of Guatemala and that many of the species will prove to be the same, seeing that Cichlasoma maculicauda of Guatemala is also found in the Chagres. Except C. maculicauda and the two species belonging to the South-American genera Acara and Geophagus, only four Cichlids have been described from Costa Rica and Panama. These are: (1) Cichlasoma citrinellum, also found in Lake Nicaragua; (2) Cichlasoma altifrons, very close to C. rodrigüem of Lake Nicaragua; (3) Cichlasoma bifasciata, allied to C. bifasciata of Lake Nicaragua; and (4) Paraneopterus stebbingi, conspecific with P. stebbingi from Southern Mexico, but also near to Neotroplus nematopterus of the Great Lakes of Nicaragua.

There are some species of Cichlids from North America, east of the Rocky Mountains and Western Eurasia. Of nearly 100 species about 80 belong to the North-American genera Ethostoma, Percina, Boleosoma, etc., comprising the dwarf Perches known as "darters." Only 6 species of these extend southward to the basin of the Rio Grande and the rivers of Chihuahua and Durango. Fossil Perches are found in the middle Eocene deposits of Wyoming.

The Centracanthidae are fresh-water fishes of North America. Of about 80 species, 8 or 9 genera, all but one, the Californian Ambloapon, are divided into 8 or 9 genera, all but one, the Californian Ambloapon...

Contributed by Smithsonian Libraries

Annotated species names
Page in PDF/image format
OCR-generated text
What we plan to do

1. Enrich the BHL with **semantic metadata** using **text mining**

2. Provide improved access to biodiversity-related digital artifacts via an **enhanced search engine** and **visualisation** of results

3. Stimulate increased collaboration, interaction and sharing of information amongst BHL users via **social media**
The grant plan

Text Mining
- Term Evolution Tracking
- Concept Extraction
- Event Extraction

Crowdsourcing

BHL Portal
- Semantic Search with Visualization
- BHL Data
- Term Inventory, Semantic Metadata

OCR Correction

Users
- Social Media
- Feedback

30 April 2014 Sophia Ananiadou
OCR error correction

• Typical errors:
  – spelling errors
  – correctly spelled but erroneous words

• Availability of BHL lexical resources: EOL and uBIO

• Correction using Google n-gram corpus
  – Challenge: Latin species names sparse in corpus

• OCR correction utility
Harvesting and tracking biodiversity terms

• Term extraction and normalisation
  – e.g., AMF = arbuscular mycorrhizai fungi with NameBankID: 10223899 (NCBI Taxonomy, Species 2000, etc.)
  – Reptile originally included reptiles (turtles, snakes, lizards) and amphibians (frogs, toads, salamanders)
  – Construction of term inventory tracking term sense evolution important for the curation of historical databases
## Extracting biodiversity entities and events

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observation</td>
<td>sightings, observations and discoveries</td>
<td>Dr. Carle Liche who <strong>claimed to have seen</strong> the man-eating tree in Madagascar in 1878 first writes of it in a letter to Dr. Omelius Fredlowski [...].</td>
</tr>
<tr>
<td>Habitation</td>
<td>geographic distribution and habitat</td>
<td>In Scotland, the Dunlin is much more <strong>abundant in summer</strong>, while in Ireland it <strong>nests</strong> in a good many localities.</td>
</tr>
<tr>
<td>Nutrition</td>
<td>food sources and trophic strategy</td>
<td>A Venus's fly trap has been seen <strong>holding fast in its grip</strong> a small frog.</td>
</tr>
<tr>
<td>Reproduction</td>
<td>pollination, courtship behaviours, mating and reproductive strategies</td>
<td>My son George observed plants of <em>O. maculata</em>, and saw many specimens of <em>Empis livadi inserting their proboscides</em> into the nectary.</td>
</tr>
</tbody>
</table>
Extracting biodiversity entities and events

• Methods:
  – automatic concept recognition (e.g., names of people, locations, species, anatomical entities, characteristics, temporal expressions)
  – automatic event extraction (observation, etc)

• Outcome:
  – automatic information extraction tools for biodiversity
  – semantic metadata added onto BHL documents
Curation, crowsourcing, collaborative annotation

Anatomical Entity Tagger

DESCRIPTION
Tags anatomical entities using Brown, UMLS and OBO Anatomy dictionary features

INPUT ANNOTATIONS
Sentence org.u_compare.shared.sentence
Chunk org.u_compare.shared.sentence
AlternateToken uk.ac.nactem.uima.eupmc
DictionaryFeatures uk.ac.nactem.uima.eupmc

OUTPUT ANNOTATIONS
NamedEntity org.u_compare.shared.semantic
Automatically recognised named entities

carrying archaeocytes,[3] and these tracks also move inert objects.[4]
It used to be claimed that glass sponges could live on nutrients dissolved in sea water and were very averse to silt.[16] However, a study in 2007 found no evidence of this and concluded that they extract bacteria and other micro-organisms from water very efficiently (about 79%) and process suspended sediment grains to extract such prey.[17] Collar bodies digest food and distribute it wrapped in vesicles that are transported by dynein "motor" molecules along bundles of microtubules that run throughout the syncytium.[3]
Sponges' cells absorb oxygen by diffusion from water into cells as water flows through body, into which carbon dioxide and other soluble waste products such as ammonia also diffuse. Archeocytes remove mineral particles that threaten to block the ostia, transport them through the mesohyl and generally dump them into the outgoing water current, although some species incorporate them into their skeletons.[3]
Carnivorous sponges

A few species that live in waters where the supply of food particles is very poor prey on crustaceans and other small animals. Most belong to the family Cladorhizidae, but a few members of the Guitaridae and Esperipodsidae are also carnivores.[18] In most cases little is known about how they actually capture prey, although some species are thought to use either sticky threads or hooked spicules.[18][19] Most carnivorous sponges live in deep waters, up to 8,840 m (5.49 mi),[20] and the development of deep-ocean exploration techniques is expected to lead to the discovery of several more.[3][18] However, one species has been found in Mediterranean caves at depths of 17–23 m (56–75 ft), alongside the more usual filter feeding sponges. The cave-dwelling predators capture crustaceans under 1 mm (0.039 in) long by entangling them with fine threads, digest them by enveloping them with further threads over the course of a few days, and then return to their normal shape; there is no evidence that they use venom.[20]
Most known carnivorous sponges have completely lost the water flow system and choanocytes. However, the genus Chondrocladia uses a highly modified water flow system to inflate balloon-like structures that are used for capturing prey.[18][21]

Endosymbionts
Linking to external dictionaries

OH-M and of residual methoxychlor. CYP3A4 does not catalyze the O-demethylation or hydroxylation of methoxychlor, but does hydroxylate mono-OH-M (ortho to the phenolic hydroxyl) (Stresser DM and Kuper D, Biochemistry 36: 2203-2210, 1997). A combination of reconstituted CYP2C19 and 3A4 in the same vessel elicits stimulation of the ortho hydroxylation of mono-OH-M compared with 2C19 alone. It is unlikely that stimulation of hydroxylation was due to protein-protein interactions, generating more active P450(s), because progression of the stimulation was time-dependent. When reconstituted CYP3A4 was added to an ongoing incubation, reformation occurred. In another experiment, we reconstituted together in vitro demonstrates that the stimulation of the phenolic methyl ether is demethylated by...
Events: associations between entities

Moreover, regional administration of L-NAME, in combination with TNF and melphalan, was studied in an isolated limb perfusion (ILP) model using BN175 soft-tissue sarcomas. Systemic treatment with L-NAME inhibited growth of adenocarcinoma significantly but was accompanied by impaired renal function.

In ILP, reduced tumour growth was observed when L-NAME was used alone. In combination with TNF or melphalan, L-NAME increased response rates significantly compared to perfusions without L-NAME (0-64% and 0-63% respectively).

An additional anti-tumour effect was demonstrated when L-NAME was added to the synergistic combination of melphalan and TNF (responses increased from 70 to 100%). Inhibition of NO synthase reduces tumour growth both after systemic and regional (ILP) treatment.

A synergistic anti-tumour effect of L-NAME is observed in combination with melphalan and/or TNF using ILP. These results indicate a possible role of L-NAME for the treatment of solid tumours in a systemic or regional setting.
Corpus annotation via crowdsourcing

• Annotated corpora needed to support development of text mining tools
• Purposeful gamification: make it fun!
• Users will be asked to identify concepts, link them to vocabularies, find associations
• Outcome:
  – large corpus with annotated concepts and events
Multi-faceted search

- Facets (dimensions) will be used to group search results semantically
- E.g., results for “nutrition” can be faceted by
  - species
  - feeding habits, food sources
  - countries

- Outcome:
  - semantic search engine
ISHER PROJECT OUTCOMES

Selected Categories event_movement/transport_role_destination:the_united_states (x)

Deadly Journeys
1212296
Deadly Journeys The death of 58 Chinese immigrants found suffocated in an airless container truck at the port of Dover, England, last month provided stark witness to the rising global traffic in people who illegally cross borders and whole continents in search of a better life. The Chinese immigrant...

Kosovo Refugee, 78, Dies on Flight to U.S.
1109015
Kosovo Refugee, 78, Dies on Flight to U.S. Mevlude Maxhuni had made it out of Kosovo, out of Macedonia and onto a jet taking her to a safe haven in the United States. But Ms. Maxhuni, 78, never set foot on American soil. She died on a Boeing 747 en route to McGuire Air Force Base in New Jersey yester...

The Wall That Keeps Illegal Workers In
1751993
The Wall That Keeps Illegal Workers In The Mexican-American border is not now and never has been out of control. The rate of undocumented migration, adjusted for population growth, to the United States has not increased in 20 years. That is, from 1980 to 2004 the annual likelihood that a Mexican wi...

Illegal Migrants Bound for U.S. Found in Box at Hong Kong Port
1240218
Illegal Migrants Bound for U.S. Found in Box at Hong Kong Port Hong Kong customs agents discovered 26 illegal Chinese migrants today, hidden in a shipping container bound for Long Beach, Calif., after they detected carbon dioxide gas emanating from the container. The container, which was scheduled t...

CUBA VOMS TO END EXODUS IN RETURN FOR A RISE IN VISAS
711264
CUBA VOMS TO END EXODUS IN RETURN FOR A RISE IN VISAS Cuba and the United States reached an agreement today under which Havana pledged to stop its citizens from fleeing the country for Florida aboard makeshift rafts and small boats. In return, Washington promised to accept at least 20,000 new Cuban ...

Some in Mexico See Border Wall As Opportunity
1764214
Some in Mexico See Border Wall As Opportunity To build, or not to build, a border of walls? The debate in the United States has started some Mexicans thinking it is not such a bad idea. Nationalist outrage and accusations of hypocrisy over the prospect have filled airwaves and front pages in Mexico...

Mexican President Thanks Bush for Support on Changes in Immigration
1764606
Mexican President Thanks Bush for Support on Changes in Immigration On the last day of his four-day visit to the United States, President Vicente Fox of Mexico spoke with President Bush to thank him for his support for the sweeping immigration changes that were passed by the Senate this week. Deputy...

Paid Notice: Deaths CASHIN, RICHARD MARSHALL
906941
Paid Notice: Deaths CASHIN, RICHARD MARSHALL, 72, international aid official, resident of New York City and Cotuit, MA, died peacefully on February 13 of complications from pulmonary fibrosis. He is survived by his wife of 46 years, Mary Walsh Cashin, four children, Anne Gol...

SALVADORANS STREAM INTO U.S., FLEETING POVERTY AND CIVIL WAR Lead: Driven by unemployment and war at home, hundreds of thousands of Salvadorans are estimated to have gone to live in the United States in recent years, according to American and Salvadoran officials. Driven by unemployment and war at ho...

In Mexico, Grim Resolve After Deaths
1256416
In Mexico, Grim Resolve After Deaths So many have lost their lives crossing into the promised land that some in Mexico were simply numb by news of the deaths of 14 more migrants abandoned in the Arizona desert. President Vicente Fox and his ministers were not. No issue has been more emotional for...
Visualisation

• Navigation of search results
• Construction of semantic similarity networks, i.e., clustering
• Outcome:
  – visualisation modules
Visualisation of results
Social digital library

• Sharing and cross-linking BHL digital objects with social media sites
• Collaborative curation: allow users to label and discuss BHL objects
• Outcome:
  – Social media integration
Evaluation

• Quantitative
  – Reliability of annotations
  – Performance of information extraction tools
  – Speed of indexing
  – Scalability

• Qualitative
  – Questionnaires
  – Focus groups
  – Evaluation workshops

• Outcome:
  – Workshops, evaluation reports