Analyzing the Biological Literature with evoText

UQAM, 1/11/2019

Charles H. Pence
@pencechp · @pencelab

UCLouvain
Institut supérieur de philosophie (ISP)

Outline

1. Tool development and future work
2. Current corpus status
3. evoText in action: analyzing biodiversity

The take-home: Could evoText be useful for your work? Get in touch!

Tool Development

Why a New Tool?

Journal articles are weird.

- They're small.
- They carry unusual metadata.
- People are used to searching/indexing them in particular ways.
- They're hard to get and come in an infinite variety of formats.

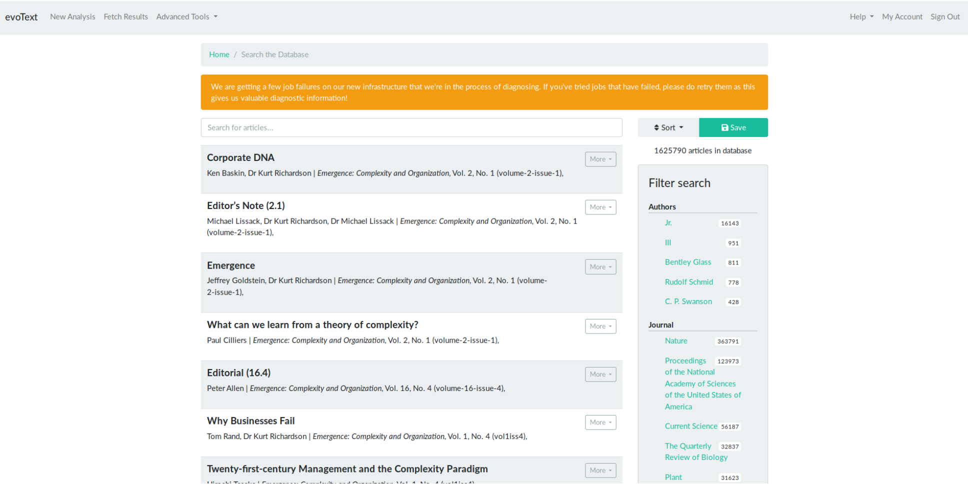
Tricks

- Have to analyze using only basic, plain text. (No guarantee we can always get anything better.)
- Ways to search/filter by journals, years (categories coming soon)
- Have to be robust to low-quality data (often entirely dependent on publishers for data quality; too many documents for manual cleaning)
- Also coming soon: citation analysis support

Current Infrastructure

Frontend: Ruby on Rails web application
Analysis backend: Ruby
Article database: Apache Solr

https://www.evotext.org/



We are getting a few job failures on our new infrastructure that we're in the process of diagnosing. If you've tried jobs that have failed, please do retry them as this gives us valuable diagnostic information!

1625790 articles in database

Corporate DNA
Ken Baskin, Dr Kurt Richardson | *Emergence: Complexity and Organization*, Vol. 2, No. 1 (volume-2-issue-1),

Editor's Note (2.1)
Michael Lissack, Dr Kurt Richardson, Dr Michael Lissack | *Emergence: Complexity and Organization*, Vol. 2, No. 1 (volume-2-issue-1),

Emergence
Jeffrey Goldstein, Dr Kurt Richardson | *Emergence: Complexity and Organization*, Vol. 2, No. 1 (volume-2-issue-1),

What can we learn from a theory of complexity?
Paul Cilliers | *Emergence: Complexity and Organization*, Vol. 2, No. 1 (volume-2-issue-1),

Editorial (16.4)
Peter Allen | *Emergence: Complexity and Organization*, Vol. 16, No. 4 (volume-16-issue-4),

Why Businesses Fail
Tom Rand, Dr Kurt Richardson | *Emergence: Complexity and Organization*, Vol. 1, No. 4 (vol1iss4),

Twenty-first-century Management and the Complexity Paradigm
Hiroshi Tasaka | *Emergence: Complexity and Organization*, Vol. 1, No. 4 (vol1iss4),

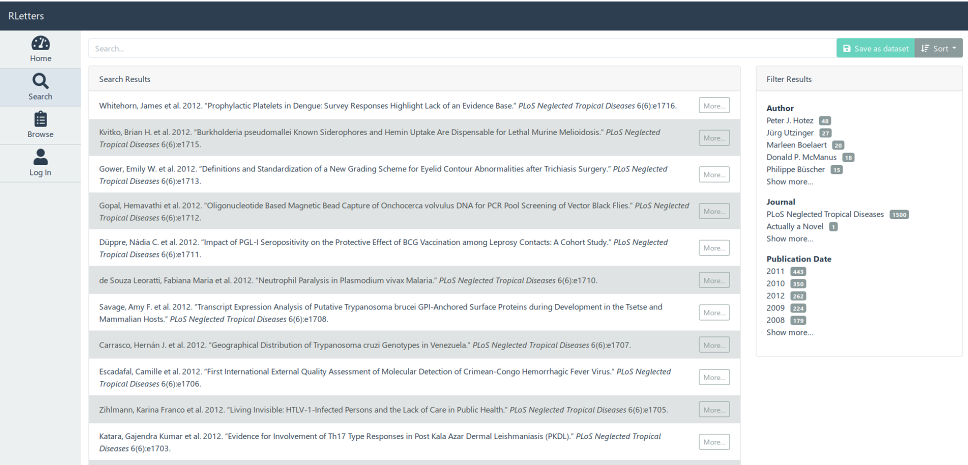
Filter search

Authors	
Jr.	16143
III	951
Bentley Glass	811
Rudolf Schmid	778
C. P. Swanson	428

Journal	
Nature	363791
Proceedings of the National Academy of Sciences of the United States of America	123973
Current Science	56187
The Quarterly Review of Biology	32837
Plant	31623

Coming soon: New version!

Frontend: Angular web app (new interface!)
Analysis backend: Go (faster analysis!)
Article database: Solr (or maybe MongoDB?)



Search Results

Whitehorn, James et al. 2012. “Prophylactic Platelets in Dengue: Survey Responses Highlight Lack of an Evidence Base.” *PLoS Neglected Tropical Diseases* 6(6):e1716.

Kvitko, Brian H. et al. 2012. “Burkholderia pseudomallei Known Siderophores and Hemin Uptake Are Dispensable for Lethal Murine Melioidosis.” *PLoS Neglected Tropical Diseases* 6(6):e1715.

Gower, Emily W. et al. 2012. “Definitions and Standardization of a New Grading Scheme for Eyelid Contour Abnormalities after Trichiasis Surgery.” *PLoS Neglected Tropical Diseases* 6(6):e1713.

Gopal, Hemavathi et al. 2012. “Oligonucleotide Based Magnetic Bead Capture of Onchocerca volvulus DNA for PCR Pool Screening of Vector Black Flies.” *PLoS Neglected Tropical Diseases* 6(6):e1712.

Düppre, Nádia C. et al. 2012. “Impact of PGL-I Seropositivity on the Protective Effect of BCG Vaccination among Leprosy Contacts: A Cohort Study.” *PLoS Neglected Tropical Diseases* 6(6):e1711.

de Souza Leoratti, Fabiana Maria et al. 2012. “Neutrophil Paralysis in Plasmodium vivax Malaria.” *PLoS Neglected Tropical Diseases* 6(6):e1710.

Savage, Amy F. et al. 2012. “Transcript Expression Analysis of Putative Trypanosoma brucei GPI-Anchored Surface Proteins during Development in the Tsetse and Mammalian Hosts.” *PLoS Neglected Tropical Diseases* 6(6):e1708.

Carrasco, Hernán J. et al. 2012. “Geographical Distribution of Trypanosoma cruzi Genotypes in Venezuela.” *PLoS Neglected Tropical Diseases* 6(6):e1707.

Escadafal, Camille et al. 2012. “First International External Quality Assessment of Molecular Detection of Crimean-Congo Hemorrhagic Fever Virus.” *PLoS Neglected Tropical Diseases* 6(6):e1706.

Zihlmann, Karina Franco et al. 2012. “Living Invisible: HTLV-1-Infected Persons and the Lack of Care in Public Health.” *PLoS Neglected Tropical Diseases* 6(6):e1705.

Katara, Gajendra Kumar et al. 2012. “Evidence for Involvement of Th17 Type Responses in Post Kala Azar Dermal Leishmaniasis (PKDL).” *PLoS Neglected Tropical Diseases* 6(6):e1703.

Filter Results

Author

Peter J. Hotez 48
Jürg Utzinger 27
Marleen Boelaert 20
Donald P. McManus 18
Philippe Büscher 15
Show more...

Journal

PLoS Neglected Tropical Diseases 1500
Actually a Novel 1
Show more...

Publication Date

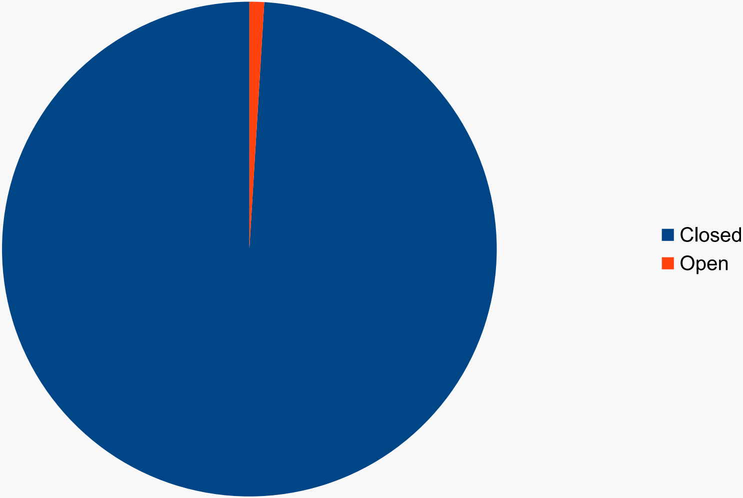
2011 443
2010 350
2012 262
2009 224
2008 179
Show more...

Current Corpus

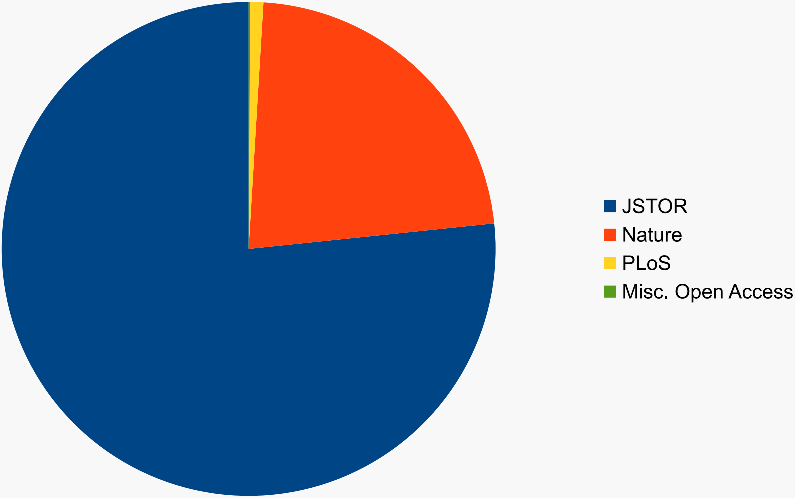
The Corpus

1,625,790 documents
334 journals

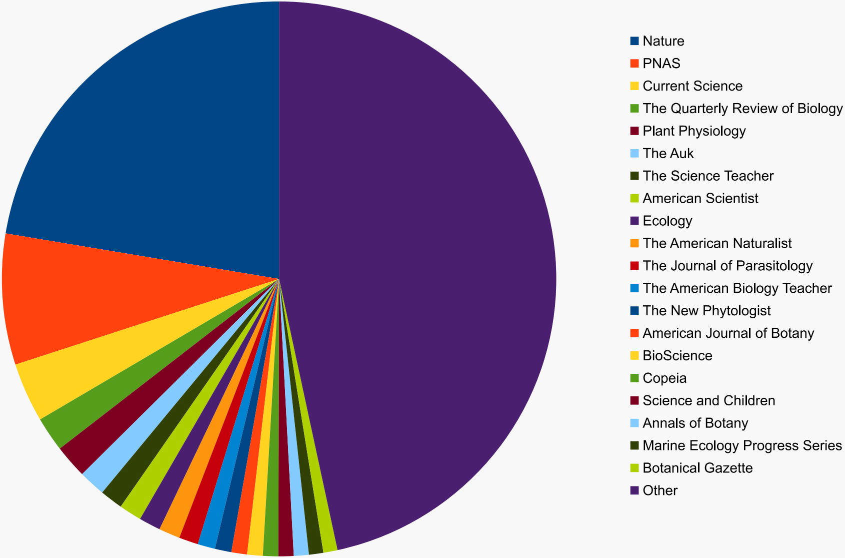
Access



Data Sources



Journals


Nature
PNAS
Current Science
The Quarterly Review of Biology
Plant Physiology
The Auk
The Science Teacher
American Scientist
Ecology
The American Naturalist
The Journal of Parasitology
The American Biology Teacher
The New Phytologist
American Journal of Botany
BioScience
Copeia
Science and Children
Annals of Botany
Marine Ecology Progress Series
Botanical Gazette
Other

Example: Biodiversity

Conceptual Analysis

Classic conceptual analysis question: What do scientists mean by **biodiversity**?

Corpus

Articles from *Conservation Biology*, from 1987–2012.

5,459 articles; 27M tokens; 427k types.

Craig-Zeta Algorithm

Take a corpus that you believe to consist of two "sub-corpora," A and B.

Goal: Find 'marker words' that distinguish A papers as opposed to B papers, and vice versa.

Craig-Zeta Algorithm

Idea: Analyze as though papers that mention biodiversity are a different sub-corpus from those that do not.

CZ Marker Words

Markers for biodiversity:

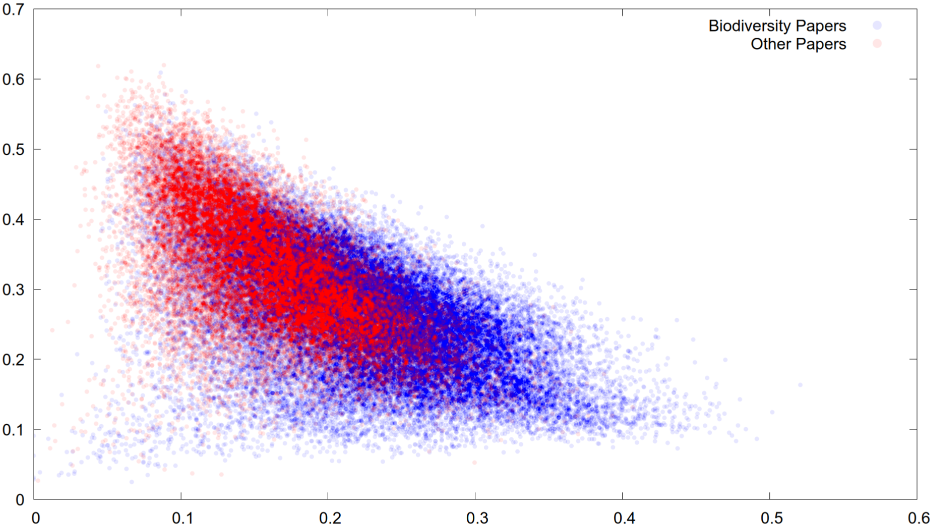
- diversity
- richness
- ecological
- protected
- *planning*
- conservation
- *development*
- *policy*
- *economic*
- assessment
- *international*
- *management*
- (Year numbers, 1997–2007)

Markers for not-biodiversity:

- population
- genetic
- breeding
- individuals
- survival
- rate
- variation
- reproductive
- mortality
- adult
- mean
- demographic
- (Year numbers, 1980–1988)

Craig-Zeta Algorithm

How do we know that the algorithm was successful?


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
Biodiversity Papers ●
Other Papers ●

Cooccurrence

What words occur (significantly often) within a given distance of ‘biodiversity’ (our focal word of interest)?

Cooccurrences (500-word window)

- biointegrity
- bioresources
- macroclimatic
- *distributive*
- *bureaucratically*
- countdown
- *neoliberalization*
- *postmodernism*
- *manifesto*
- cataloguing
- underprotected
- biopiracy
- hotspots
- coextinctions

A first suggestion…

Biodiversity is unusually related with words indicating **social and political context.**

A first suggestion…

Biodiversity is unusually related with words indicating **social and political context.**

Good! This lines up both with what historians of science and the practitioners themselves have told us about the self-image of the biodiversity paradigm.

Comparing Definitions

Could we try to look for places where **different** definitions of biodiversity have been used in different places in the literature?

Comparing Definitions

Could we try to look for places where **different** definitions of biodiversity have been used in different places in the literature?

One idea: Biodiversity is an inherently **spatial** concept. What parts of the globe has it been applied to in different contexts?

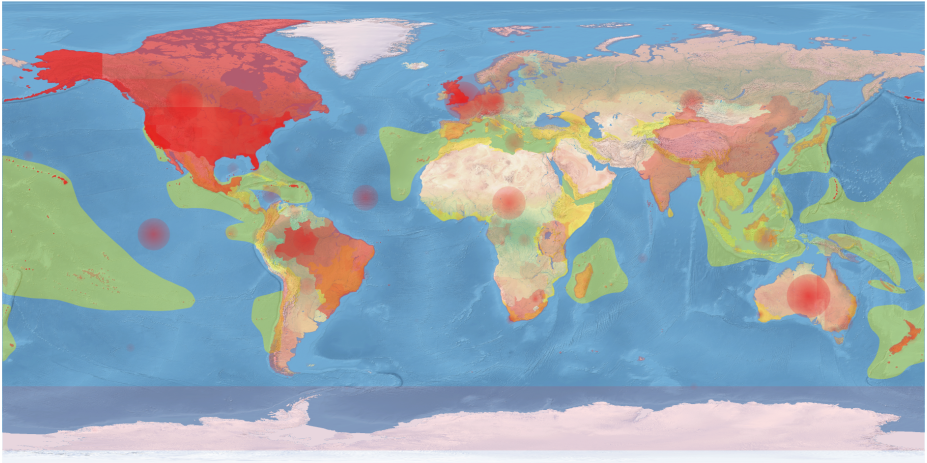
Named Entity Recognition

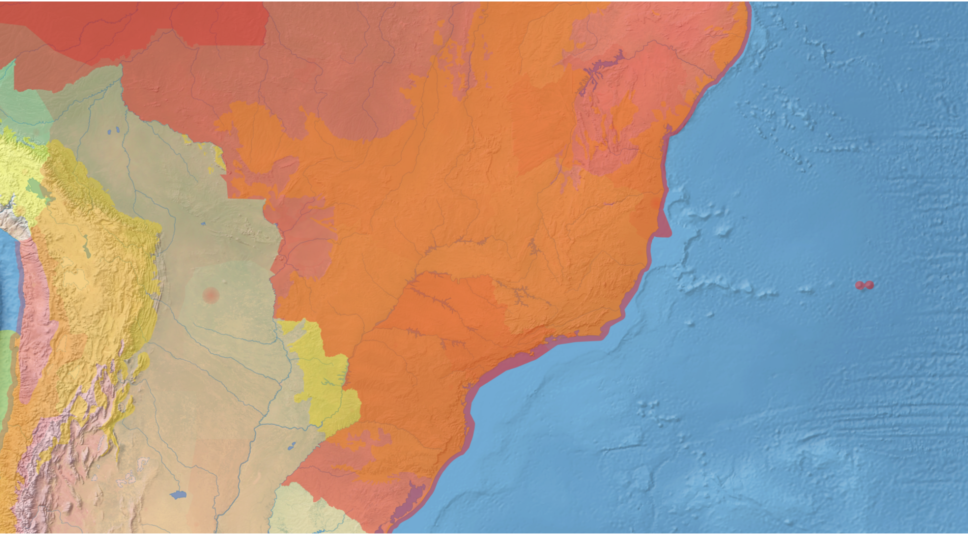
Look for any instances of proper place names throughout the *Conservation Biology* corpus, map them.

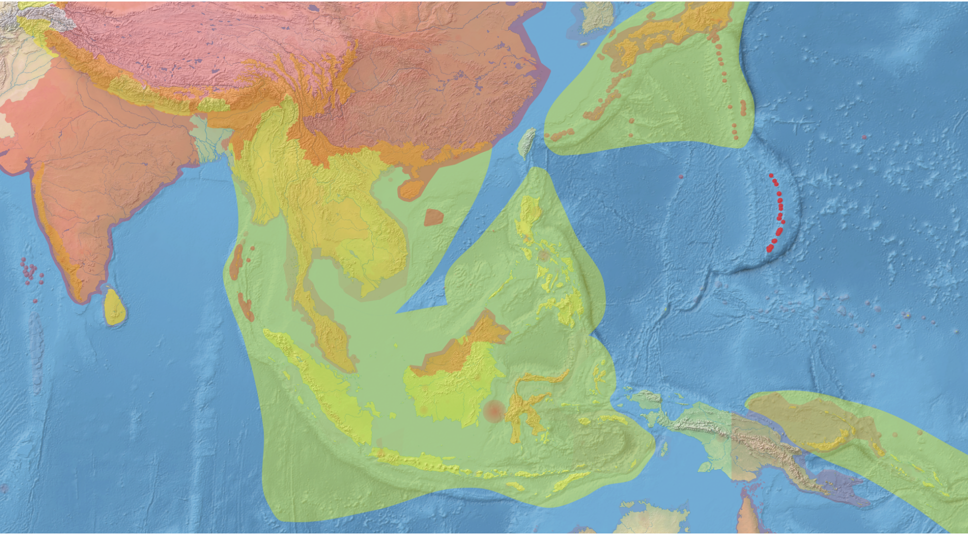
Named Entity Recognition

Look for any instances of proper place names throughout the *Conservation Biology* corpus, map them.

Unsolved problem for evoText: Geocoding! How to get from “Australia” to a set of geographic coordinates on a map? Currently using 3rd-party API (with a large free service tier).







A second suggestion…

Various definitions of biodiversity seem to **come apart** when we compare scientific and political uses of the concept.

Back to evoText

Other tools available:

- General word frequency analysis (1-gram or N-gram, with tf-idf scores, configurable block creation behavior, etc.)
- Collocation analysis (like cooccurrence but for direct bigrams)
- Some simple graphing-by-date
- Export to various reference manager citation formats

Questions?

charles@charlespence.net
https://pencelab.be
@pencechp · @pencelab

The Pence Lab

Wallonie - Bruxelles
International.be

fnrs
LA LIBERTÉ DE CHERCHER

Craig-Zeta Algorithm

1. Divide corpus A and corpus B into blocks of 500 words.
2. For each word, determine how many blocks in which that word appears. (Discard any words that appear in *every* block.)
3. For each word, compute the Zeta score:

$$Z_w = \frac{\mathrm{count}_A}{n_A} + \frac{n_B - \mathrm{count}_B}{n_B} \tag{1}$$

That is: the fraction of blocks of A that contain w plus the fraction of blocks of B that *do not* contain w. Z_w values range from 2 (appears in every block of A and no blocks of B) to 0 (appears in no blocks of A and every block of B).

Collocation Algorithm

1. Inputs: word of interest t, window s.
2. Divide the corpus into n blocks of size s.
3. For every word w in the corpus, compute both f_w (how many blocks contain w) and f_{wt} (how many blocks contain both w and t).
4. Compute a score for the significance of the wt pair. In this case, we used mutual information:

$$I_{wt} = \log\left(\frac{f_{wt} \cdot n}{f_t \cdot f_w}\right) \tag{2}$$