



ELSEVIER

# Text Analytics at Elsevier

Antony Scerri

Elsevier Labs

2012-06-20

# Introduction

- Elsevier Labs
  - Who we are and what we do
- Primary use of GATE
  - Proof-of-concept development
- SciVerse and text mining
  - Exposing text mining as a service on all our content
- Elsevier Strategy
  - Semantic Enrichment
  - Big Data Analytics

# PoC Work

- GATE at a glance
  - easy to use and not a closed box
  - plug in new processors, customise existing ones
  - JAPE lets you analyse your content
- Visual inspection of results
- Plenty of options
  - Comparison of components, e.g. different POS taggers

# Example Projects

- PharmaEffects
- SD Trends
- BrainLink

# PharmaEffects

- No known vocabulary list at the time
- Easy to find trigger phrases
- Build a list
  - Find trigger phrases with gazetteer
  - Extract surrounding terms using JAPE
  - Collect matched phrases (with freq.)
  - Split into valid / invalid
  - Repeat process filtering known examples
  - Use valid examples to find other trigger phrases

# SD Trends

- 2008
- Marine geology
- 10K articles
- 163 concept classes
- Initially built using ClearForest tool
- Comparison of functionality to GATE
  - DIAL4 vs JAPE
  - ANNIE tagger vs CF tagger
- Overall near equal functionality

# SD Trends



Trends in Marine Geology

Brought to you by Elsevier Labs

Concepts ?

- Anthropogenic Measures/Policies
- Events
  - Climate events
  - Climate periods
    - Abstract climate periods
    - Named climate periods
  - Meteorologic (weather) events
- Forcing Mechanisms
  - Anthropogenic forcing mechanisms
  - Climate forcing mechanisms
  - Meteorologic mechanisms
- Geological Settings
- Locations
- Research Tools

Tag Cloud | Graph | Relational | C2C | Help

air pollution atlantic multidecadal oscillation deforestation desertification **el niño** eutrophication **forestry**

greenhouse gas emissions **la niña** land management **land use** north atlantic oscillation reforestation

Date periods ?  Include all Showing : 2007 Q4

Occurrence Threshold ? : >0 >3

Chosen concepts ?

Events -> Climate events

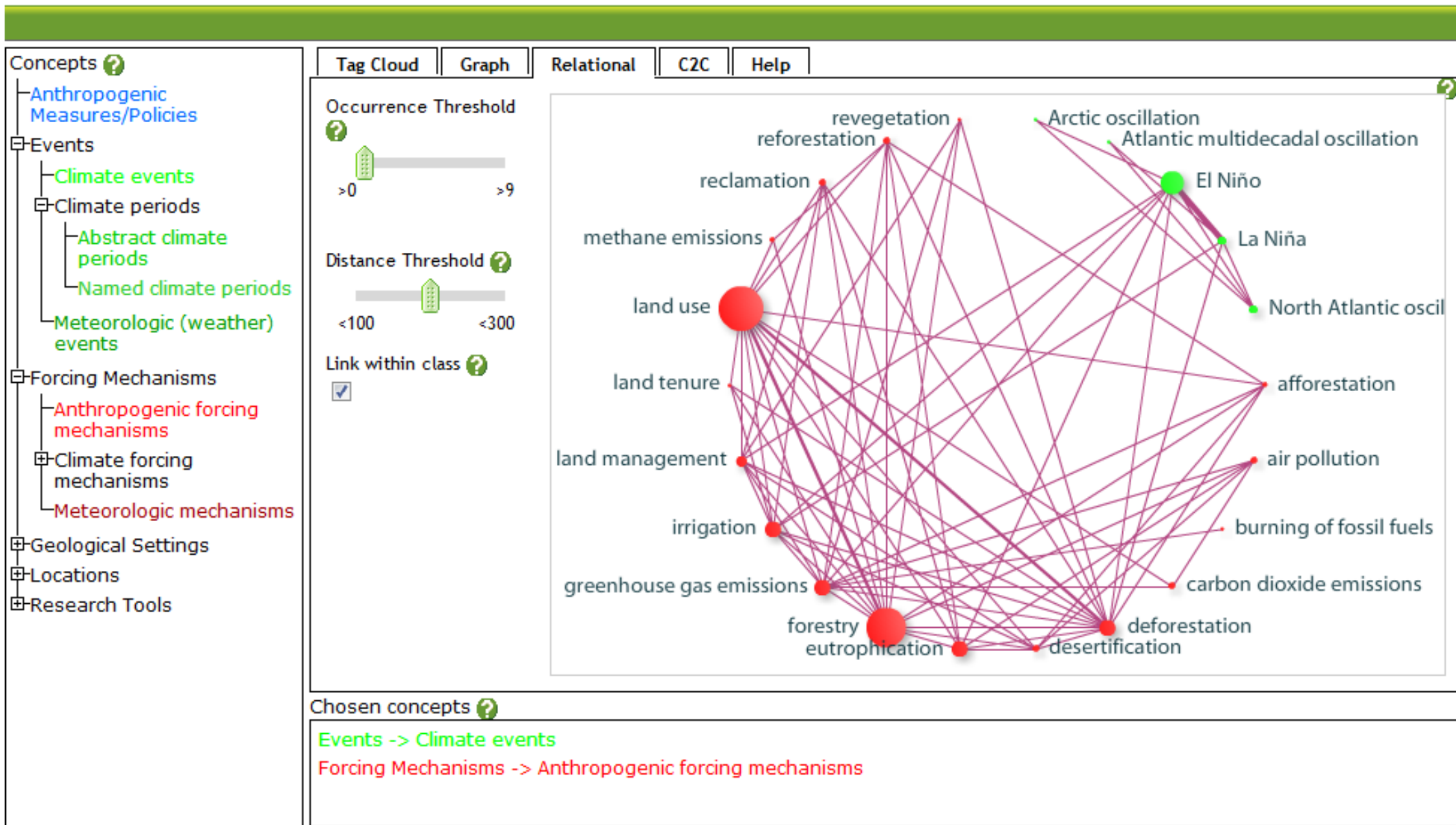
Forcing Mechanisms -> Anthropogenic forcing mechanisms

# SD Trends



Trends in Marine Geology

Brought to you by Elsevier Labs





# BrainLink

- Link from journals to brain atlas
- Neuroanatomical feature identification
- Species identification
- Recall over precision
- Gazetteer

# BrainLink

- Compile lexicon
  - Collect ontology / taxonomy sources
  - Extract labels
    - XQuery over XML and SPARQL over RDF
  - Map concepts
  - “munging” labels
    - Identify synonymous words – region, area, part etc
    - Unpack notation – A, region B of = region B of A
      - Hippocampus, alveus of the = alveus of the hippocampus
    - Inserting subphrases – A, B = B of the A
      - stria terminalis, lateral division = lateral division of the stria terminalis

# BrainLink cont.

- Neuroanatomical structures
  - 5.5K concepts
  - 7 ontologies
- Gazetteer term list
  - 124K terms
  - Preparation takes time
  - Common tasks for cleanup or expansion
- Recall over precision
  - Missing subpart of a structure
  - Semantic ambiguity e.g. hippocampus

# BrainLink Demo

<http://www.sciencedirect.com/science/article/pii/S0006899311023237>

The screenshot shows a ScienceDirect article page. At the top, there are navigation links for SciVerse, ScienceDirect, Hub, ScienceDirect, Scopus, and Applications. The user is logged in as 'antony scerri' and can click 'Logout' or 'Go to SciVal Suite'. The article title is 'Expression and localization of myosin-1d in the developing nervous system' by Andrew E. Benesh, Jonathan T. Fleming, Chin Chiang, Bruce D. Carter, and Matthew J. Tyska. The article is from 'Brain Research', Volume 1440, 27 February 2012, Pages 9–22. The article is a Research Report. The abstract discusses the expression and localization of myosin-1d in the developing nervous system. A sidebar on the right, titled 'BrainLink Powered by Brain Navigator', provides information about brain structures mentioned in the article. The sidebar includes a search bar and a table of structures mentioned in the article.

Home + Recent Actions | Browse | Search | My settings | My alerts | Help

Back to results | « Previous 4 of 8 Next » | Export citation | PDF (4542 K) | More options... | Search ScienceDirect | Search

**Brain Research**  
Volume 1440, 27 February 2012, Pages 9–22

Research Report

## Expression and localization of myosin-1d in the developing nervous system

Andrew E. Benesh<sup>a, c, 1</sup>, Jonathan T. Fleming<sup>a, c</sup>, Chin Chiang<sup>a, c</sup>, Bruce D. Carter<sup>b, c</sup>, Matthew J. Tyska<sup>a, c</sup>

<sup>a</sup> Department of Cell and Developmental Biology, Vanderbilt University Medical Center, Nashville, TN, USA  
<sup>b</sup> Department of Biochemistry, Vanderbilt University Medical Center, Nashville, TN, USA  
<sup>c</sup> Program in Developmental Biology, Vanderbilt University Medical Center, Nashville, TN, USA

Accepted 26 December 2011. Available online 8 January 2012.

<http://dx.doi.org/10.1016/j.brainres.2011.12.054>, How to Cite or Link Using DOI | Cited by in Scopus (0)

Permissions & Reprints

### Abstract

Myosin-1d is a monomeric actin-based motor found in a wide range of tissues, but highly expressed in the nervous system. Previous microarray studies suggest that myosin-1d is found in oligodendrocytes where transcripts are upregulated during the maturation of these cells. Myosin-1d was also identified as a component of myelin-containing subcellular fractions in proteomic studies and mutations in MYO1D have been linked to autism. Despite the potential implications of these previous studies, there is little information on the expression and localization of myosin-1d in the developing nervous system. Therefore, we analyzed myosin-1d expression patterns in the peripheral and central nervous systems during postnatal development. In mouse sciatic nerve, myosin-1d is expressed along the axon and in the ensheathing myelin compartment. Analysis of mouse cerebellum prior to myelination at day 3 reveals that myosin-1d is

### BrainLink Powered by Brain Navigator

Information about brain structures in this article

Select a Structure

Structure	Mentioned	Images
<a href="#">brain stem</a>	1	<a href="#">Yes</a>
<a href="#">brainstem</a>	1	<a href="#">Yes</a>
<a href="#">central nervous systems</a>	3	<a href="#">Yes</a>
<a href="#">cerebellar nuclei</a>	4	
<a href="#">cerebellum</a>	17	<a href="#">Yes</a>
<a href="#">cerebral cortex</a>	2	<a href="#">Yes</a>
<a href="#">cerebrum</a>	2	<a href="#">Yes</a>
<a href="#">hippocampal formation</a>	1	<a href="#">Yes</a>
<a href="#">spinal cord</a>	2	
<a href="#">thalamus</a>	1	<a href="#">Yes</a>

### Related articles

- Developmental expression of neuronal nitric ...  
*Brain Research*
- Developmental expression of neuronal nitric ...  
*Brain Research*
- Comparison of RPTP $\zeta$ / $\beta$ , phosphacan, and trk...  
*Molecular Brain Research*
- Murine numb regulates granule cell maturatio...  
*Developmental Biology*
- Autoradiographic localization of inhibitory ...  
*Brain Research*

[View more related articles](#)

# BrainLink Demo

Search ScienceDirect  Search

**BrainLink** Powered by Brain Navigator  
Information about brain structures in this article ?

cerebellum (17) ▾

[Back to List](#)

[Search in ScienceDirect](#) | [View in Brain Navigator](#)  
[2D Images](#) | [3D Images](#) | [Mentioned In Article \(17\)](#)

[Mouse](#) | [Rat](#) | [Monkey](#)

The Mouse Brain in Stereotaxic Coordinates  
by K Franklin and G Paxinos

Bregma -6.0 mm      Interaural -2.2 mm




Figure 81

Search ScienceDirect  Search

**BrainLink** Powered by Brain Navigator  
Information about brain structures in this article ?

cerebellum (17) ▾

[Back to List](#)

[Search in ScienceDirect](#) | [View in Brain Navigator](#)  
[2D Images](#) | [3D Images](#) | [Mentioned In Article \(17\)](#)

- 1. Abstract**  
...ensheathing myelin compartment. Analysis of mouse **cerebellum** prior to myelination at day 3 reveals that myosin...
- 2. Introduction**  
... rat cerebral cortex, spinal cord, brainstem, and **cerebellum**, in addition to a number of other tissues (Bahler...
- 3. Introduction**  
...ms (CNS). In the CNS, our analysis focused on the **cerebellum**, where Myo1d expression is limited to neurons, ex...
- 4. Myo1d exhibits a developmentally regulated distribution in the cerebellum**

# Custom Gazetteer

- Standalone and embedded in GATE
- Features
  - Lucene tokeniser stream
  - Longest only (removing any inner matches)
  - Multi-threaded
  - Lower memory footprint
  - Word synonym sets
  - Optional backing by Lucene index



# Other projects

- AnswerBot
  - Semantic search in Lucene
  - 1M terms, 690 docs @ 113M took <3mins
- HCIR – Information Availability
  - Query Analytics Workbench
  - Clustering around NPs
- Domain independent statements in abstracts
  - identify rhetorical structure like problems, results, methods etc
  - JAPE to extract common patterns

# Challenges

- Complexity of XML DTD's
  - Handling markup
  - Aligning result to original XML
- Scientific literature
  - General lack of tagged corpus (esp. large body)
  - Availability of models trained against such corpus
    - GENIA better than most
      - Java port
    - Being able to compare with GATE
  - Breadth of domains we cover



# Summary

- Quick and easy PoC setup
- Comparing/evaluating component options
  - POS taggers
  - parser
- Gazetteer
  - Identify known concepts
- JAPE
  - find candidates (NEs and expressions)

# SciVerse and text mining

- Opening product platform
  - Provide access to our content and services
  - Integration of 3<sup>rd</sup> party “gadgets”
  - Integration of additional services
- Open Miner Service
  - Start to provide additional services alongside our content
- Example applications
  - Lipids (24K terms)
  - Glossary (65K terms)

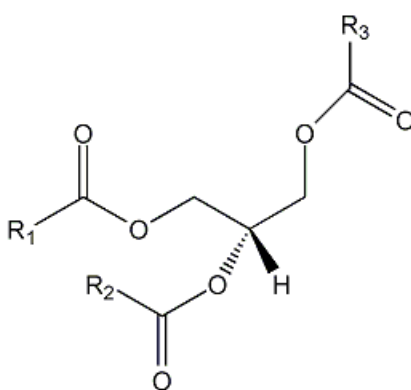
# Lipids App

Export citation PDF (1185 K) More options... Search ScienceDirect Search

Lipid Structures (beta)

## LIPID MAPS Structure database (LMSD)

Common Name: TAG



LM ID: LMGL03010000  
Common Name: TAG  
Systematic Name: triacyl-sn-glycerol  
Synonyms: Triacylglycerol  
Exact Mass: -  
Formula: -  
Category: [Glycerolipids \[GL\]](#)  
Main Class: [Triacylglycerols \[GL03\]](#)  
Sub Class: [Triacylglycerols \[GL0301\]](#)  
PubChem Substance ID (SID): [24701469](#)  
METABOLOMICS ID: -  
KEGG ID: [C00422](#)  
HMDB ID: [HMDB05357](#)  
CHEBI ID: -  
InChIKey: -  
InChI: -  
Status: Active

Download file MDLMOL

Structure viewing options: [Static Image](#) [MarvinView Applet](#) [JmolApplet](#) [ChemDraw](#) (?)

3.6. Cell-based studies of ADRP and lipoprotein metabolism LSD-2 99

javascript:void(0);

# Glossary App

tings | My alerts Help

Glossary Search

<b>Atrophy</b> [ Cerebral Metabolism, Brain Imaging in Encyclopedia of Stress (Second Edition) ]
Reduction in cell mass.
<b>Atrophy</b> [ Space, Health Risks of in Encyclopedia of Stress (Second Edition) ]
A diminution or degeneration of bodily tissues or organs.
<b>Atrophy</b> [ Diffusion MRI in Neurological Disorders in Diffusion MRI ]
Partial or complete wasting away of a part of the body or organ or tissue.
<b>Atrophy</b> [ Glossary in Principles and Practice of Implant Dentistry ]
1. A wasting away. 2. A diminution in the size of a cell, tissue, organ, or part.
<b>Atrophy:</b> [ Progressive Central Nervous System Disorders in Physical Rehabilitation ]
Wasting or loss of muscle tissue resulting from disease or lack of use.
<b>Atrophy</b> [ Glossary in Reptile Medicine and Surgery (Second Edition) ]
The physiologic or pathologic reduction in size of a cell, tissue, organ, or region in the body.
<b>Atrophy</b> [ Glossary of some of the terms used by therapists (most terms are described as used in the chapters) in Finnie's Handling the Young Child with Cerebral Palsy at Home (Fourth Edition) ]
The wasting of muscles
<b>Atrophy</b> [ Glossary in Gait Analysis (Fourth Edition) ]
loss of bulk, especially of a muscle.
<b>Atrophy</b> [ Glossary in The Dog Breeder's Guide to Successful Breeding and Health Management ]
Wasting away
<b>Atrophy</b> [ Glossary in Cyclura ]
Wasting of tissue or an organ.

ts with diffuse disease systemic therapy is indicated Useful approaches include therapy with

# Custom Gazetteer cont.

- Lazy options 😊
  - Punctuation and diacritic sensitivity
  - Comma transposition eg A, B = B A
  - Stemming
- Parallel tokenization of content to accommodate patterns
  - De-hyphenation
  - Parentheses
    - Dropping contents
    - Extracting contents

# Elsevier Strategy

- Semantic Enrichment
- Big Data Analytics

# Semantic Enrichment

- Adding value to existing content
- Elsevier's Linked Data repository
  - RDF store and API's for search and retrieval

# Big Data Analytics

- Project - Content Analytics Toolkit (CAT)
- Statistical NLP
- High-level use cases:
  - Operate on whole or parts of document text
    - Images and other media will come later
  - Clustering
  - Classification
  - Training of:
    - POS tagger, parser, NE taggers
  - Relation extraction



# CAT – High-level overview

- Pre-processing pipeline
  - NLP tooling independent
  - Simple model to cope with bare minimum but handle any additional data
- Data storage
  - Support querying at various levels of granularity
  - Provide minimal statistics over data
- Statistical NLP
  - Generating features
  - Applying appropriate algorithms

# CAT – Pre-Processing

- Unpack XML
- Normalise characters and whitespace
- Minor modification to content to ease processing
- Tokenise and POS tag
  - Plus noun and verb phrase chunking
- Additional steps will likely include NE identification

# CAT - Architecture

- Amazon Cloud based
- S3 source repository
- Mongo pre-process results storage
  - GridFS
  - \*SON records
- Column store?
- Storm deploying jobs across machines

# CAT – Statistical NLP

- Just starting out
- Repeatable process to do:
  - Sample selection
  - Feature engineering
  - Vectorizing
  - Model training and evaluation
- All at scale
- Gaining experience

# CAT – Next Steps

- Evaluating tools : Mahout, Mallet, ClearTK, Weka, Vowpal Wabbit...
- Evaluate POS taggers
  - Manual tag a corpus
  - Train our own tagger

# Courses

- Couseira (Stanford)
  - Machine Learning
  - NLP
  - Probabilistic Graphical Models



ELSEVIER

# Thanks

Antony Scerri

[a.scerri@elsevier.com](mailto:a.scerri@elsevier.com)