

Text mining workflows for indexing archives with automatically extracted semantic metadata

Riza Batista-Navarro¹, Axel Soto¹,
William Ulate² and Sophia Ananiadou¹
¹University of Manchester
²Missouri Botanical Garden

Outline (1)

- Introduction
 - challenges in information discovery/search
 - semantic metadata generation as a named entity recognition (NER) task
- The Argo text mining workbench
 - system overview and features
 - workflow construction, configuration and execution
 - visual inspection of generated annotations

Outline (2)

- Constructing NER workflows for generating semantic metadata
 - medical history archives
 - biodiversity legacy literature
- Exploring search indexes containing semantic metadata
 - introduction
 - overview of Elasticsearch
 - query examples

Outline (3)

- Example applications
 - Disambiguation in the History of Medicine search system
 - Biodiversity Heritage Library query expansion
- Conclusions



Biodiversity Heritage Library

- <http://www.biodiversitylibrary.org/>
- a consortium of botanical and natural history libraries
- stores digitised legacy literature on biodiversity
- currently holds 180,000 volumes = 50+ million pages (PDFs and OCR-generated text)
- open-access

BHL's keyword-based search and browsing

[About BHL](#)[Help](#)

Inspiring discovery through free access to biodiversity knowledge.

The Biodiversity Heritage Library works collaboratively to make biodiversity literature openly available to the world as part of a global biodiversity community.

BHL also serves as the foundational literature component of the Encyclopedia of Life (EOL).



Search across books and journals, scientific names, authors and subjects

[ADVANCED SEARCH](#)

Browse Our Collection By:

[Titles](#)[Authors](#)[Date](#)[Collection](#)

SUPPORT

Help Support BHL

BHL's existence depends on the financial support of its patrons. Help us keep this free resource alive!

[Donate Now](#)

BHL's advanced search functionality ⁷ (also keyword-based)

Books/JournalsArticles/ChaptersAuthorsSubjectsScientific Names

Title:

Author Last Name:

Volume:

Edition:

Year (YYYY):

Subject:

Language:

(Any Language)

Collection:

(Any Collection)

Search

SUPPORT



Help Support BHL

BHL's existence depends on the support of its patrons. Help us keep this free resource alive!

Donate Now

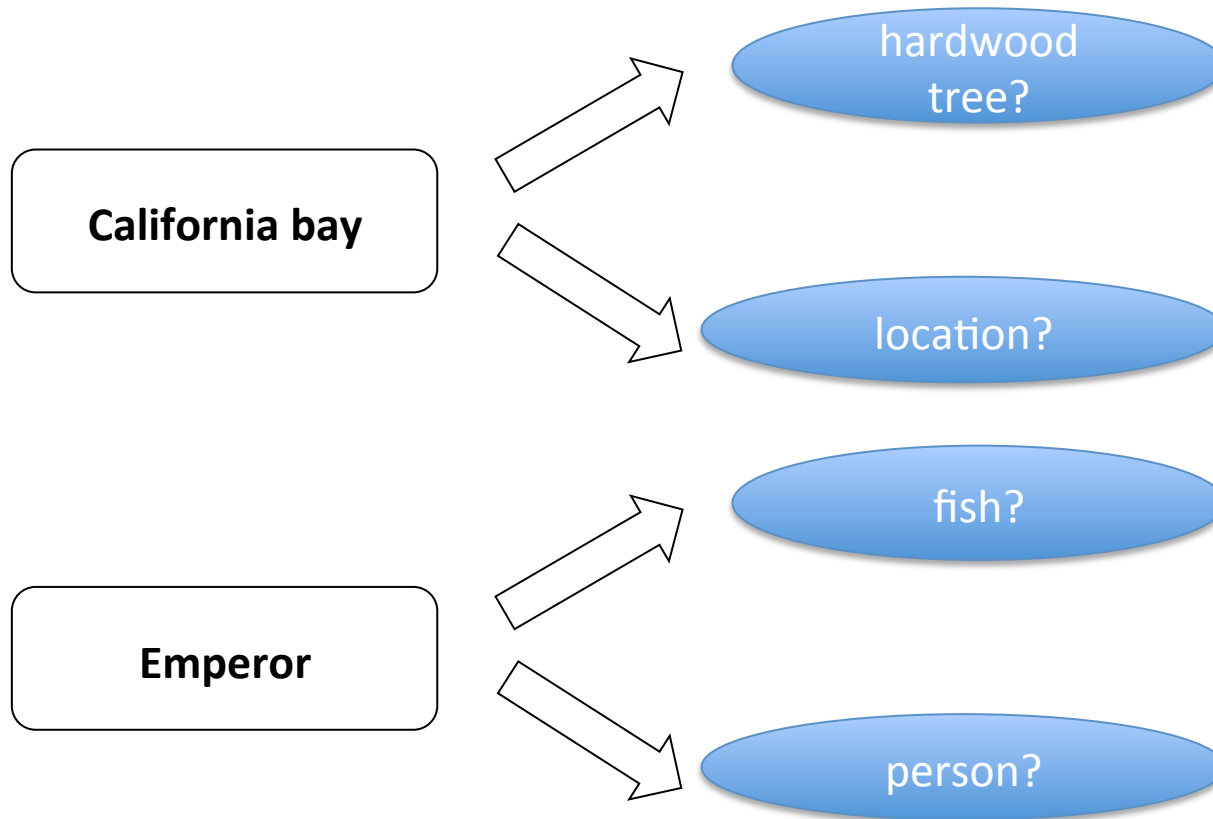
Featured Collection

Celebrating Alfred Russel Wallace



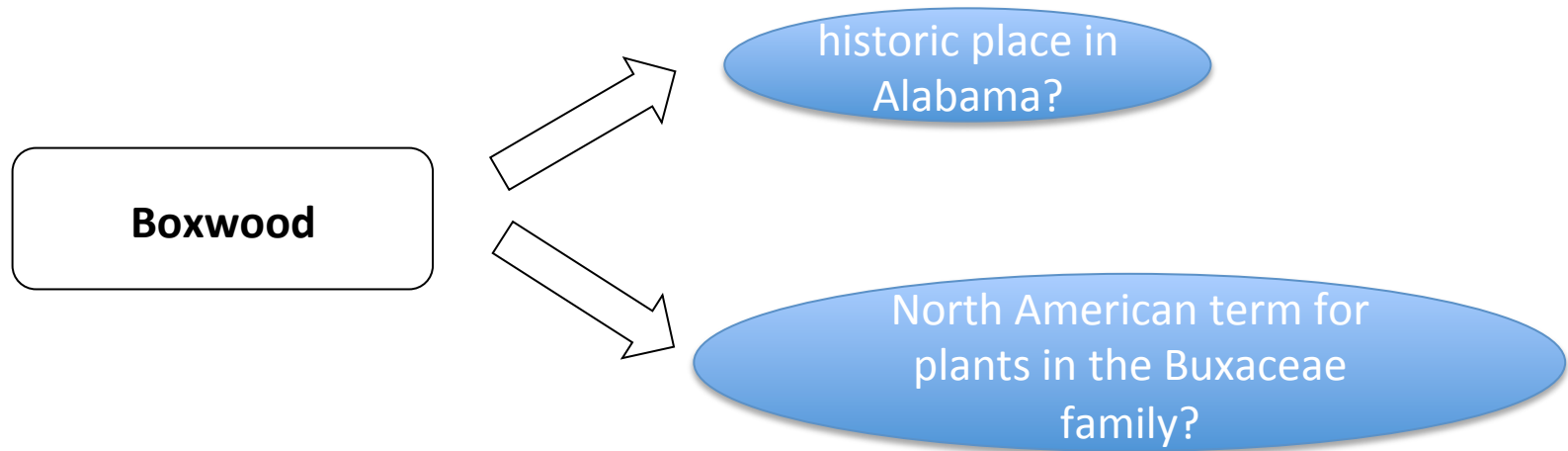
What's wrong with keyword-based search?

(1) Inability to disambiguate



What's wrong with keyword-based search?

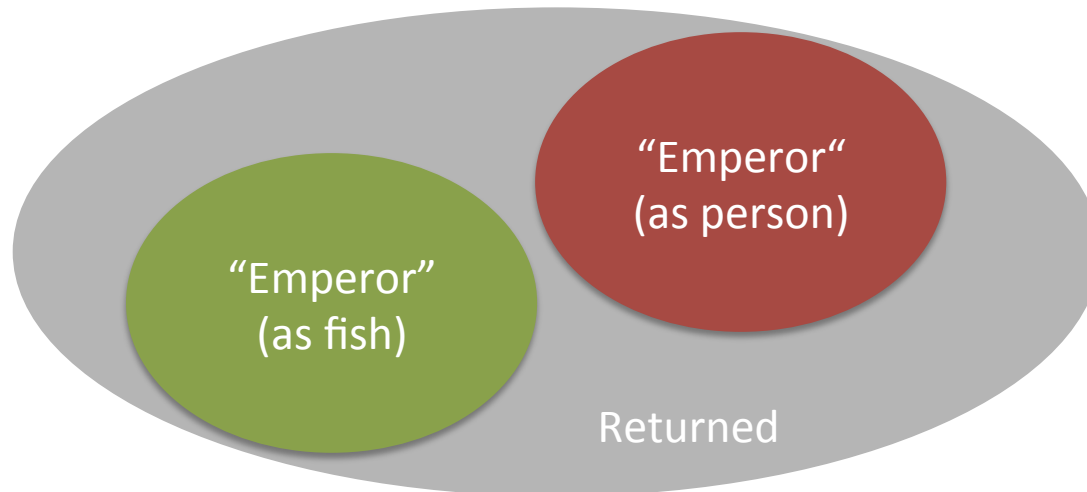
(1) Inability to disambiguate



What's wrong with keyword-based search?

(1) Inability to disambiguate

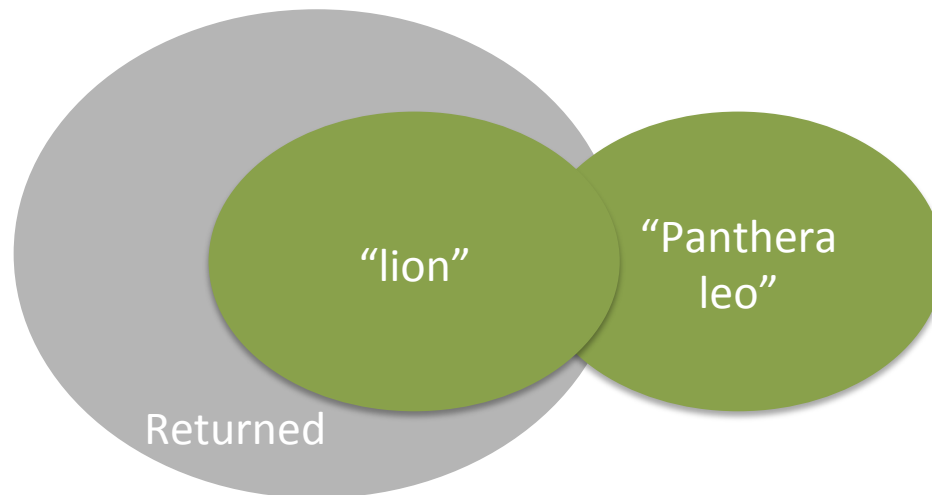
- Implications:
 - less precise search results
 - even documents irrelevant to what you have in mind are returned



What's wrong with keyword-based search?

(2) Inability to account for variants

- Implications:
 - limited coverage
 - information overlook



Solution: Semantic metadata generation¹² using named entity recognition (NER)

- task of automatically demarcating mentions
 - detecting their boundaries (e.g., character offsets)
 - placing them into predefined categories

Tax
Platymantis banahao-See list of holotype and paratypes for that species.
Acknowledgments

Early field work by **Per** E. H. Taylor, **Per** A. C. Alcala, and **Per** W. C. Brown provided most of the collections and much of the data on the ecology and zoogeography of this species-group.

Recent field work by **Per** A. C. Alcala, **Per** A. C. Diesmos, **Per** A. Ross.

Per E. L. Alcala, and numerous people who assisted them have provided data on populations occurring on **Location** southern Luzon and associated small islands.

We are indebted to **Per** R. C. Drewes (CAS), **Per** J. E. Cadle (MCZ), **Per** R. F. Inger (FMNH), **Per** P. C. Gonzales (PNM), and **Per** E. N. Arnold and **Per** B. T. Clarke (BMNH) for permission to examine material in their collections.

We also wish to thank **Per** A. E. Leviton, **Per** R. C. Drewes, and **Per** R. F. Inger for their critiques during the preparation of this paper.

Drawings were prepared by **Per** C. Sudekum, **Org** California Academy of Sciences.

Named entity recognition (NER)

- cast as a sequence labelling task
 - sequence = tokens in a sentence
- approaches
 - dictionary-based
 - rule-based
 - machine learning (ML)-based
 - hybrid

Dictionary-based NER

- **Sample entries in a gazetteer:**
- **Sample text**

Ho Chi Minh	PROVINCE	The
Ho Chi Minh City	CITY	final
...		five
Johannesburg	CITY	include
Johannesburg	PROVINCE	Mexico
...		City
Mexico	PROVINCE	,
Mexico Beach	CITY	Riyadh
Mexico City	CITY	,
Mexico Crossing	CITY	Johannesburg
...		,
Riyadh	CITY	Ho
...		Chi
Tehran	CITY	Minh
Tehran	PROVINCE	City

Dictionary-based NER

- **Sample entries in a gazetteer:**
- **Sample text matched (in BIO)**

Ho Chi Minh	PROVINCE
Ho Chi Minh City	CITY
...	
Johannesburg	CITY
Johannesburg	PROVINCE
...	
Mexico	PROVINCE
Mexico Beach	CITY
Mexico City	CITY
Mexico Crossing	CITY
...	
Riyadh	CITY
...	
Tehran	CITY
Tehran	PROVINCE

The	O	O
final	O	O
five	O	O
include	O	O
Mexico	B-CITY	B-PROVINCE
City	I-CITY	O
,	O	O
Riyadh	B-CITY	O
,	O	O
Johannesburg	B-CITY	B-PROVINCE
,	O	O
Ho	B-CITY	B-PROVINCE
Chi	I-CITY	I-PROVINCE
Minh	I-CITY	I-PROVINCE
City	I-CITY	O

Dictionary-based NER

✓ Advantages

- simple
- many readily available dictionaries/lexica

✗ Disadvantages

- dictionaries can become too big
- yet, none of them complete or comprehensive enough
- overlaps between categories, e.g., many people and places have the same names

Rule-based NER

- Regular expressions
 - checking for capitalisation
 - checking for numbers
- Function words for extracting, e.g., locations
 - Capitalized word + {city, centre, river} indicates location
Examples: ***New York city, Hudson river***
 - Capitalized word + {street, boulevard, avenue} indicates location
Examples: ***Fifth avenue***

Rule-based NER

- Context patterns
 - [PERSON] earned [MONEY]
Example: *John earned £20*
 - [PERSON] joined [ORGANISATION]
Example: *Sam joined IBM*
 - [PERSON], the [JOBTITLE]
Example: *Mary, the teacher*

Rule-based NER

- still not so simple:
[*PERSON/ORGANISATION*] fly to [*LOCATION*]
Examples: *Jerry flew to Japan*
Delta flies to Europe
Birds fly to the nest
- match patterns defined in a gazetteer
 - dictionary of person names:
[John, Jerry, Mary, Frank, David, ...]
Jerry is a person's name but not *Delta* nor *Birds*.

Rule-based NER

✓ Advantages

- handcrafted rules can be very precise
- only small amount of development data needed

✗ Disadvantages

- domain-dependent
- expensive development and test cycle

Shortcomings of dictionary- and rule-based approaches

- Failure to generalise
 - first word of a sentence is also usually capitalised
 - multiword expressions
- Inability to disambiguate
 - Jordan the **person** vs. Jordan the **location**
 - JFK the **person** vs. JFK the **airport**
 - May the **person** vs. May the **month**

Shortcomings of dictionary- and rule-based approaches

- Upkeep/maintenance
 - No gazetteer contains all existing proper names
 - New proper names constantly emerge
 - products, brands
 - scientific discoveries (e.g., planets, stars, medicines)
 - Multiple variants can emerge for the same entity
 - John Smith
 - J. Smith
 - Prof. J Smith

ML-based approaches to NER

- Supervised learning
 - labelled training examples
 - methods
 - hidden Markov models (HMMs)
 - naïve Bayes
 - decision trees
 - support vector machines (SVMs)
 - conditional random fields (CRFs)

ML-based approaches to NER

- Semi-supervised learning
 - small percentage of training examples is labelled, the rest is unlabelled
 - methods
 - bootstrapping
 - active learning
 - co-training
 - self-training
- Unsupervised learning
 - labels must be automatically discovered
 - method: clustering

Conditional random fields (CRFs)

- a widely used algorithm for sequence labelling
- finds the most probable label sequence \mathbf{y} given an observation sequence \mathbf{x}

$$\mathbf{y} = \operatorname{argmax}_{\mathbf{y}} p_{\lambda}(\mathbf{y}|\mathbf{x})$$

where \mathbf{x} consists of the sequence of tokens from input text

Conditional random fields (CRFs)

- computation of probability

$$p_{\lambda}(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_{\mathbf{x}}} \cdot \exp\left(\sum_{i=1}^n \sum_{j=1}^m \lambda_j f_j(y_{i-1}, y_i, \mathbf{x}, i)\right)$$

Diagram illustrating the components of the CRF probability formula:

- normalisation factor**: Points to $\frac{1}{Z_{\mathbf{x}}}$
- weight**: Points to λ_j
- feature function**: Points to $f_j(y_{i-1}, y_i, \mathbf{x}, i)$
- summation over all tokens**: Points to the inner summation $\sum_{j=1}^m$
- summation over all feature functions**: Points to the outer summation $\sum_{i=1}^n$

- feature function: characterises the input

$$f_i(x, y) = \begin{cases} 1, & \text{if 1}^{\text{st}} \text{ letter of } x \text{ is uppercase} \\ 0, & \text{otherwise} \end{cases}$$

Conditional random fields (CRFs):

Feature types

27

- character n -grams (e.g., 2, 3, 4-grams)
- lexical and contextual
 - current word, lemma, part-of-speech (POS) tag
 - word n -grams: around W_0 in $[-3, \dots, +3]$ window
- suffixes and prefixes (e.g., with lengths 2 to 4)

Conditional random fields (CRFs):

Feature types

- orthographic

initial-caps

all-caps

lonely-initial

all-digits

contains-dots

punctuation-mark

single-char

contains-hyphen

- semantic

- matches between tokens and names in gazetteers or controlled vocabularies

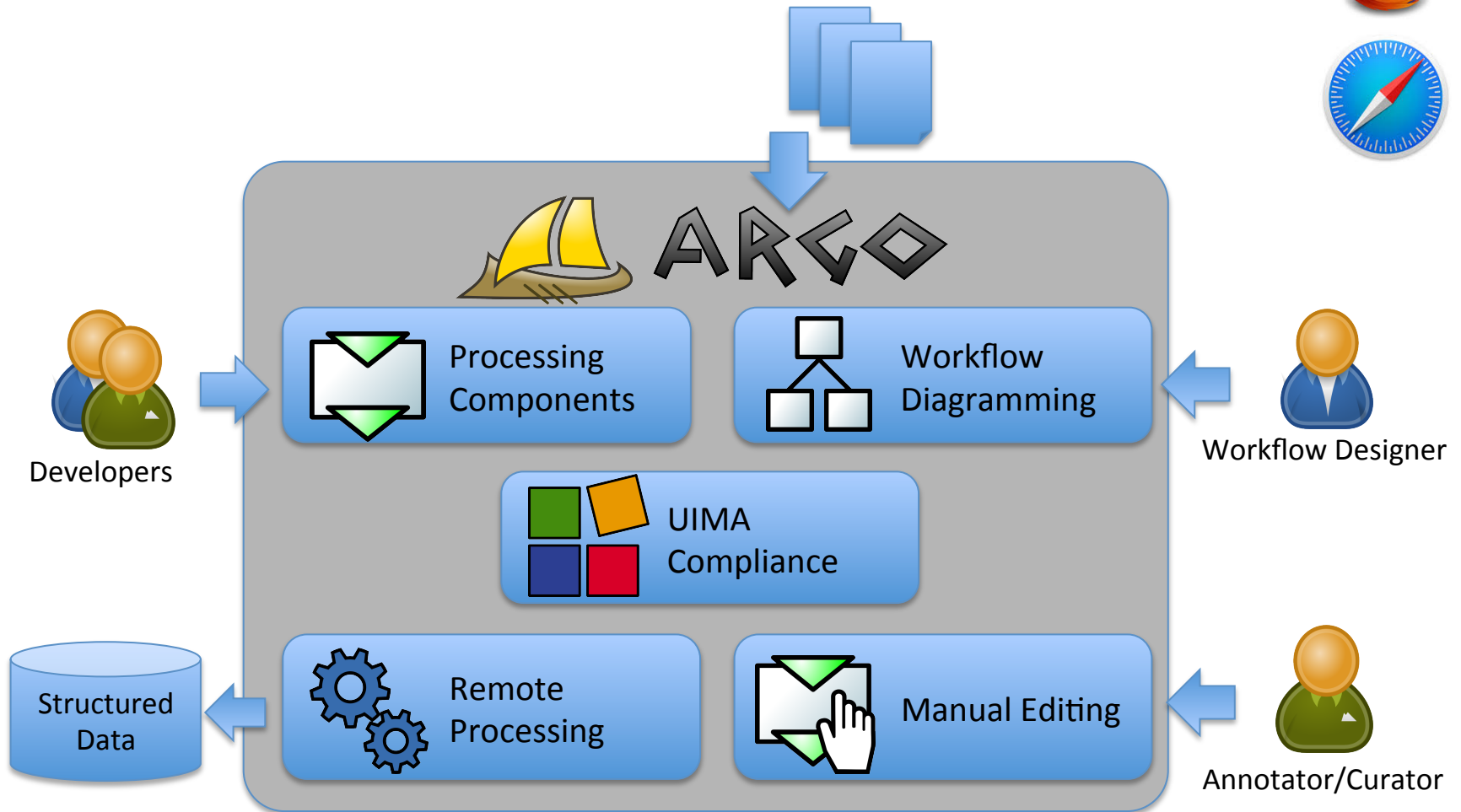
Pipelining various tools for NER

- Sentence splitting
 - to define a sequence
- Tokenisation
 - to generate the basic unit of analysis, i.e., tokens
- Lemmatisation, POS-tagging
 - to generate lexical and contextual features
- Gazetteer matching
 - to generate semantic features

Questions so far?

Argo: a generic text mining workbench

(<http://argo.nactem.ac.uk>)



Workflows

riza.batista@manchester.ac.uk ▼

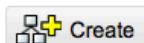


Workflows

Processes

Documents

Workflows



Create



Edit



Run



Delete



Share



Details



COPD BioNLPST Corpus generation



BioNLP



Tokenisation + POS Tagging



Copy of Demo: Chemical Entity Recogniser (ChER)



Twitter Test



Demo: Chemical Entity Recogniser (ChER)



Example: Metabolic processes - Manual annotation



Example: Metabolic processes - Manual correction



Example: Metabolic processes - Automatic annotation



Example: TRY ME FIRST!



Demo: Chemical Entity Recogniser (ChER)



DESCRIPTION

A sample workflow demonstrating an optimal solution to the chemical entity recognition task, as described in Batista-Navarro et al, 2015 (<http://www.jcheminf.com/qc/content/7/S1/S6>).

Content, in this case the PubMed abstract with PMID 23411304, is read in as plain-text.

In order to run this workflow, make sure it is selected before clicking on the "Run" button. You'll be able to observe the progress of processing in the "Processes" tab. This tab will also give you access to the Manual Annotation Editor that you can use to view the produced annotations.

Note that certain components may take up to a minute to initialise due to their loading statistical models and other supporting resources.

Processes



riza.batista@manchester.ac.uk ▼

Workflows

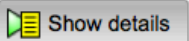
Processes

Documents

Processes



Delete



Show details

▶ 05 Sep 2016 23:47:15	Running - 62% Complete
▶ 27 May 2016 10:36:20	Finished
▶ 17 May 2016 20:25:20	Finished
▶ 10 Mar 2016 00:55:19	Finished
▶ 10 Mar 2016 00:52:51	Finished
▶ 10 Mar 2016 00:52:00	Finished
▶ 10 Mar 2016 00:50:53	Finished
▶ 10 Mar 2016 00:49:57	Finished
▶ 25 Feb 2016 20:45:48	Finished
▶ 25 Feb 2016 20:34:18	Error
▶ 25 Feb 2016 10:16:49	Finished
▶ 15 Feb 2016 21:16:39	Finished
▶ 15 Feb 2016 20:54:47	Error

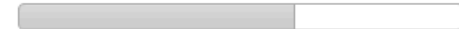
▶ Unnamed process

STARTED

05 Sep 2016 23:47:15

STATUS

Running



LAUNCH INTERACTIVE COMPONENTS

[Launch Manual Annotation Editor](#)

Documents



riza.batista@manchester.ac.uk ▼

Workflows

Processes

Documents

Documents

Refresh

Create

Upload

- [-] Argo
 - [-] BioNLP Shared Task
 - [+] CG
 - [-] PC
 - development
 - training
 - [-] My Documents
 - COPD-AutoAnnotatio
 - COPD-BioNLPST
 - COPD-NoRelations
 - [+] COPD-PhenomeNet
 - COPD-practice
 - COPD-Raw
 - COPD-TMAssisted
 - COPD-TMRelations

- LICENSE
- PMID-10072383.a1
- PMID-10072383.a2
- PMID-10072383.txt
- PMID-10329687.a1
- PMID-10329687.a2
- PMID-10329687.txt
- PMID-10424772.a1
- PMID-10424772.a2
- PMID-10424772.txt
- PMID-10462544.a1
- PMID-10462544.a2
- PMID-10462544.txt

Components

- Readers
 - loads corpora/document collections
 - provide support for various formats, e.g., plain text, XML, TSV, stand-off
- Analytics
 - natural language processing tools
 - tokenisers, POS taggers, parsers, named entity recognisers
- Consumers
 - serialisation to files (e.g., XML, TSV) and databases

Configuration

Arrange ▼

Edit Properties

More ▼

All changes saved

Readers

- BioC Reader
- BioC Web Service Reader
- BioCreative CHEMDNER Reader
- BioNLP ST Data Reader
- Document Reader
- EUPMC Reader
- ElsevierReader
- Input Text Reader
- Kleio Search
- PK DDI Corpus Provider
- PubMed Abstract Reader
- RDF Reader
- RDF Web Service Reader
- XMI Reader
- XMI Web Service Reader
- XML Reader

Analytics

- Anatomical Entity Tagger
- Annotation Remover
- Brat BioNLP ST Comparator
- Brown Dictionary Feature Extractor
- GNERF & Dictionary Matcher

```

graph TD
    A[Kleio Search] --> B[GENIA Sentence Splitter]
    B --> C[GENIA Tagger]
    C --> D[Manual Annotation Editor]
    D --> E[XMI Writer]
    D --> F[RDF Writer]
    D --> G[BioC Writer]
    
```

GENIA Tagger

DESCRIPTION

Tags biological named entities: proteins, cell lines, cell types, DNAs, and RNAs. It has its own tokeniser, part-of-speech tagger, and shallow parser. The models were trained on the GENIA corpus.

Project website:
<http://www.nactem.ac.uk/GENIA/tagger/>

INPUT ANNOTATIONS

Sentence	org.u_compare.shared.syntactic
Token	org.u_compare.shared.syntactic

OUTPUT ANNOTATIONS

GeniaToken	jp.ac.u_tokyo.s.is.www_tsujii.tools
Chunk	jp.ac.u_tokyo.s.is.www_tsujii.tools
RNA	org.u_compare.shared.semantic.k
Protein	org.u_compare.shared.semantic.k
DNA	org.u_compare.shared.semantic.k
CellType	org.u_compare.shared.semantic.k
CellLine	org.u_compare.shared.semantic.k

Configuration

The screenshot displays the Argo workflow configuration interface. On the left, a sidebar lists various components under 'Readers' and 'Analytics'. The 'Readers' section includes BioC Reader, BioC Web Service, BioCreative CHEM, BioNLP ST Data R, Document Reader, EUPMC Reader, ElsevierReader, Input Text Reader, Kleio Search, PK DDI Corpus Pro, PubMed Abstract R, RDF Reader, RDF Web Service R, XMI Reader, XMI Web Service R, and XML Reader. The 'Analytics' section includes Anatomical Entity T, Annotation Remover, Brat BioNLP ST Comparator, Brown Dictionary Feature Extr, and CHEBI Soft Dictionary Matche.

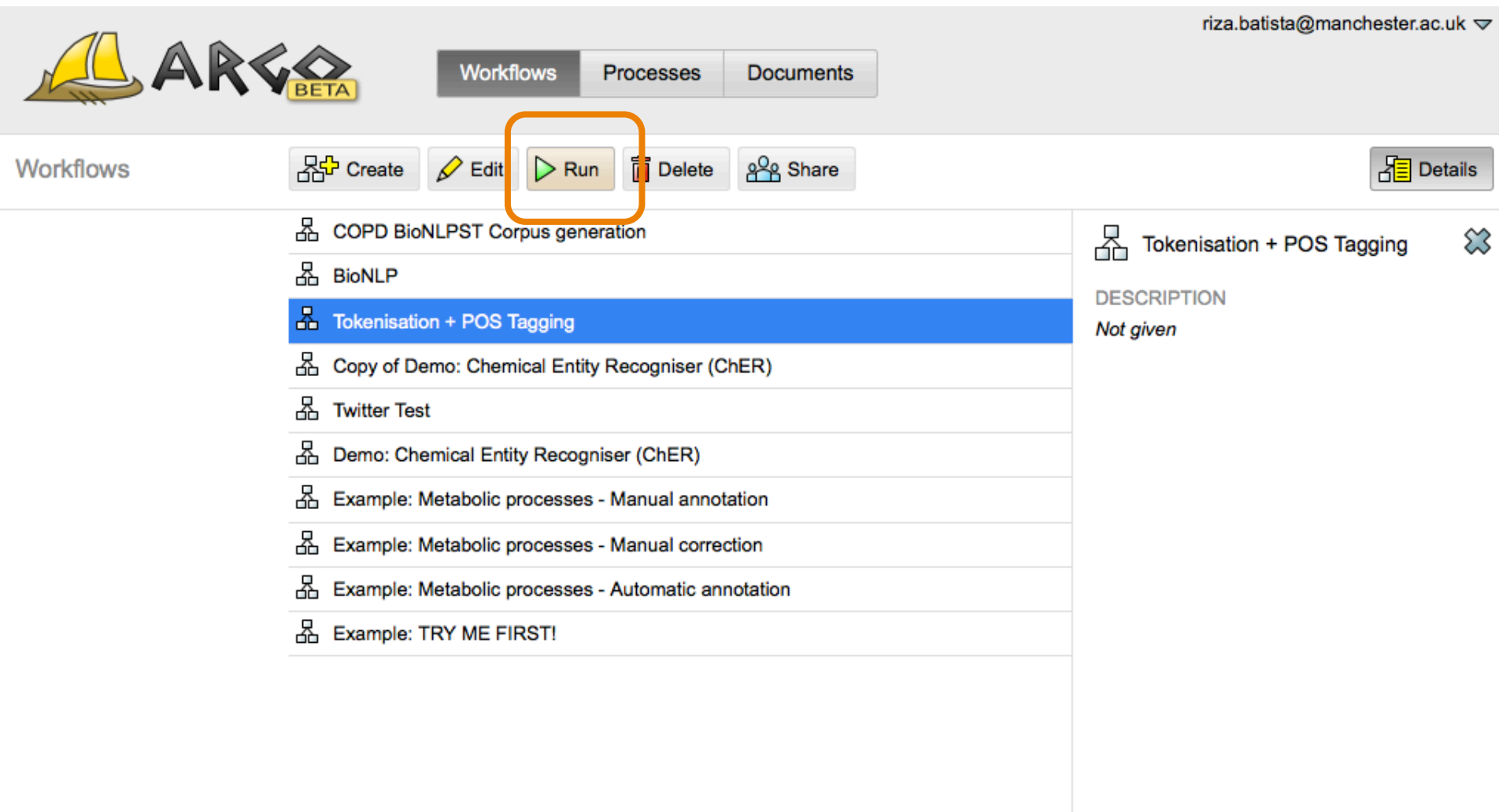
The main area shows a 'Settings: GENIA Tagger' dialog box. The 'Parameters' tab is active, displaying three settings:

Parameter	Value
tokenize	True
chunkTag	True
neTag	False

The dialog box also has a 'Visual settings' tab and buttons for 'OK', 'Save to File...', 'Load from File...', and 'Cancel'.

On the right side of the interface, there is a list of entities and their associated URIs. The entities listed are 'Cell type' and 'CellLine', both with the URI 'org.u_compare.shared.semantic.t'.

Execution



ARGO BETA

riza.batista@manchester.ac.uk ▼

Workflows Processes Documents

Workflows

Create Edit **Run** Delete Share


Details

- COPD BioNLPST Corpus generation
- BioNLP
- Tokenisation + POS Tagging**
- Copy of Demo: Chemical Entity Recogniser (ChER)
- Twitter Test
- Demo: Chemical Entity Recogniser (ChER)
- Example: Metabolic processes - Manual annotation
- Example: Metabolic processes - Manual correction
- Example: Metabolic processes - Automatic annotation
- Example: TRY ME FIRST!

Tokenisation + POS Tagging ✕

DESCRIPTION
Not given

Monitoring



WorkflowsProcessesDocuments

riza.batista@manchester.ac.uk ▾

Processes

Delete

Show details

▶ 05 Sep 2016 23:47:15	Running - 62% Complete
▶ 27 May 2016 10:36:20	Finished
▶ 17 May 2016 20:25:20	Finished
▶ 10 Mar 2016 00:55:19	Finished
▶ 10 Mar 2016 00:52:51	Finished
▶ 10 Mar 2016 00:52:00	Finished
▶ 10 Mar 2016 00:50:53	Finished
▶ 10 Mar 2016 00:49:57	Finished
▶ 25 Feb 2016 20:45:48	Finished
▶ 25 Feb 2016 20:34:18	Error
▶ 25 Feb 2016 10:16:49	Finished
▶ 15 Feb 2016 21:16:39	Finished

▶ Unnamed process ✕

STARTED
05 Sep 2016 23:47:15

STATUS
Running

LAUNCH INTERACTIVE COMPONENTS

Launch Manual Annotation Editor

Visual inspection of results: the Manual Annotation Editor

Create

Change Label

Move

Delete

Remove from Index

Finish Editing

The family **Tomoceridae** includes 149 species in 16 genera, grouped in two subfamilies, **Tomocerinae** with 131 species and **Lepidophorellinae** with 18 species (Bellinger et al. 2010). **Tomocerinae** are distributed across the whole Holarctic region, extending locally as south as the mountains of Northern Sumatra. They are conspicuous by their large size and abundance in forest litter, but are also diversified and frequent in the caves of different regions of Europe, eastern Asia and North America, with about 30 troglobitic species. Many of these cave species have a reduced number of eyes and reduced pigment (Christiansen 1964). However, few species exhibit strong morphological adaptation to cave life. The most remarkable species in this respect is **Tritomurus falcifer** Cassagnau, 1958, which apparently is limited to a small karst of the central Pyrénées. In the present paper, we describe from a Croatian cave a second highly troglomorphic species, **Tritomurus** veles sp. n., already recorded as **Tomoceridae** gen. sp. in Lukić and Deharveng (2008). We also introduce several new morphological characters for the taxonomy of **Tomoceridae**, discuss the validity of the genus **Tritomurus** and comment the world distribution of reduced-eyed **Tomoceridae**.

Annotations

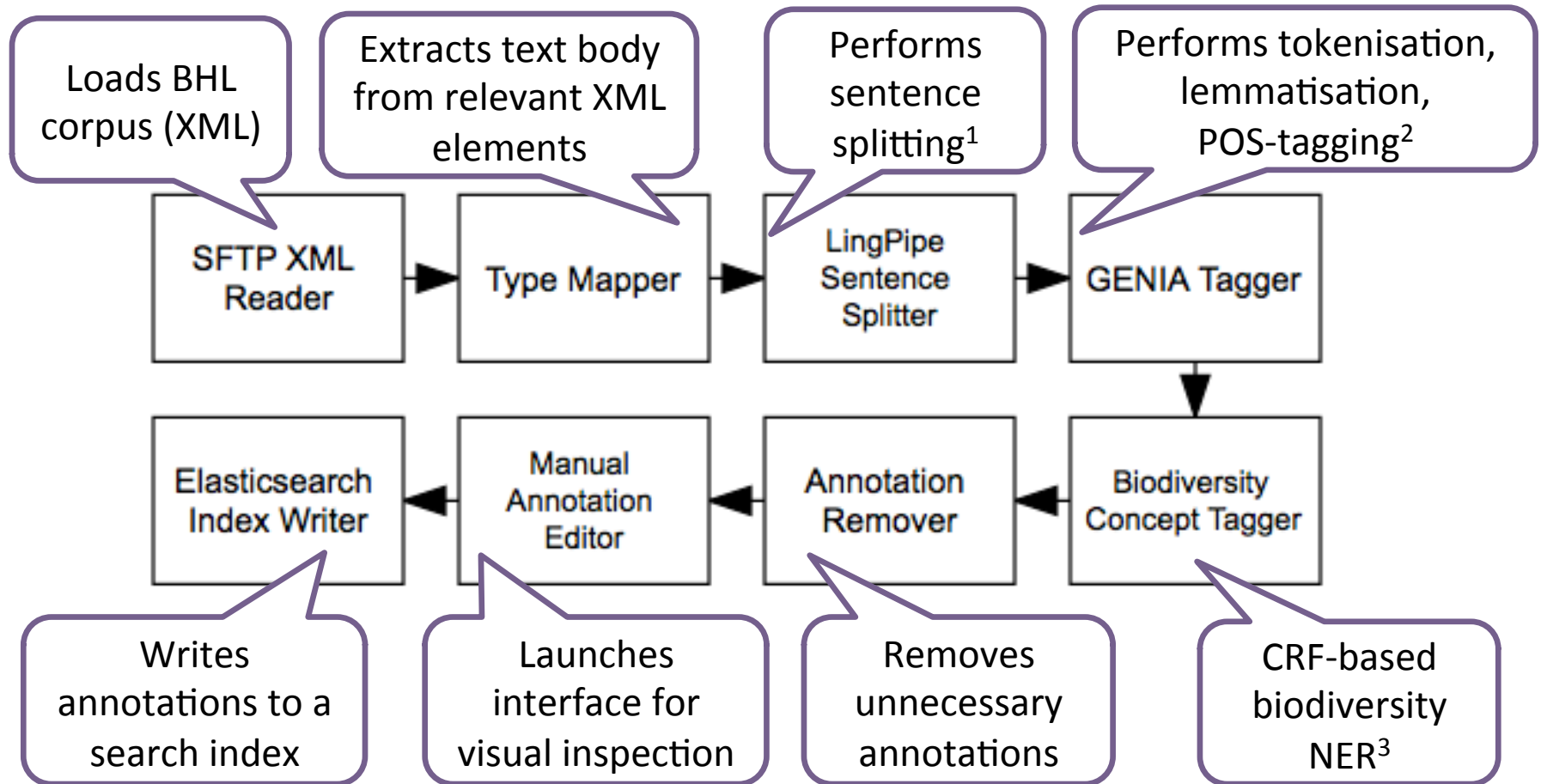
Labels

Show

☒ Labels in document
 ☐ All labels

- ☒ Taxon
- ☒ TemporalExpression
- ☒ Person
- ☒ Location
- ☒ Trait
- ☐ GeniaToken
- ☒ Species
- ☐ Chunk

Generating semantic metadata with NER workflows: biodiversity literature




¹LingPipe: <http://alias-i.com/lingpipe>

²GENIA Tagger: <http://www.nactem.ac.uk/GENIA/tagger>

³NERsuite: <http://nersuite.nlplab.org>

Generating semantic metadata with NER workflows: biodiversity literature



19838748.xml
19872366.xml
19899910.xml
19900099.xml
19917603.xml
1994458.xml
19971771.xml
19994429.xml
20002782.xml
2005011.xml
2005521.xml
20064524.xml
20079272.xml
2008001.xml
2015444.xml
20161542.xml
20161908.xml
20171977.xml
2017855.xml
20200025.xml

Create

to the **apex** of the rostrum ; in front of these elevations the surface gradually slopes downwards. The cardiac area is distinctly circumscribed. The branchial area is of considerable extent and crossed transversely by a V-shaped impression, one limb of which passes to the posterior part of the cervical groove, while the other reaches the outer boundary of the cardiac area. The ocular peduncles are **short** and **stout**, with the corneee deeply **pigmented** ; the antennal flagellum is almost twice the length of the **carapace**. The pterygostomial area possesses a series of **well-marked elevated** lines

The ischium of the external maxillipedes has the inner margin broadly **rounded**, and the outer and distal border prolonged into a **subacute lobe**, the external surface is cro

Finish Editing

Annotations Labels

Show ☒ Labels in document ☐ All labels

☒ **AnatomicalEntity**

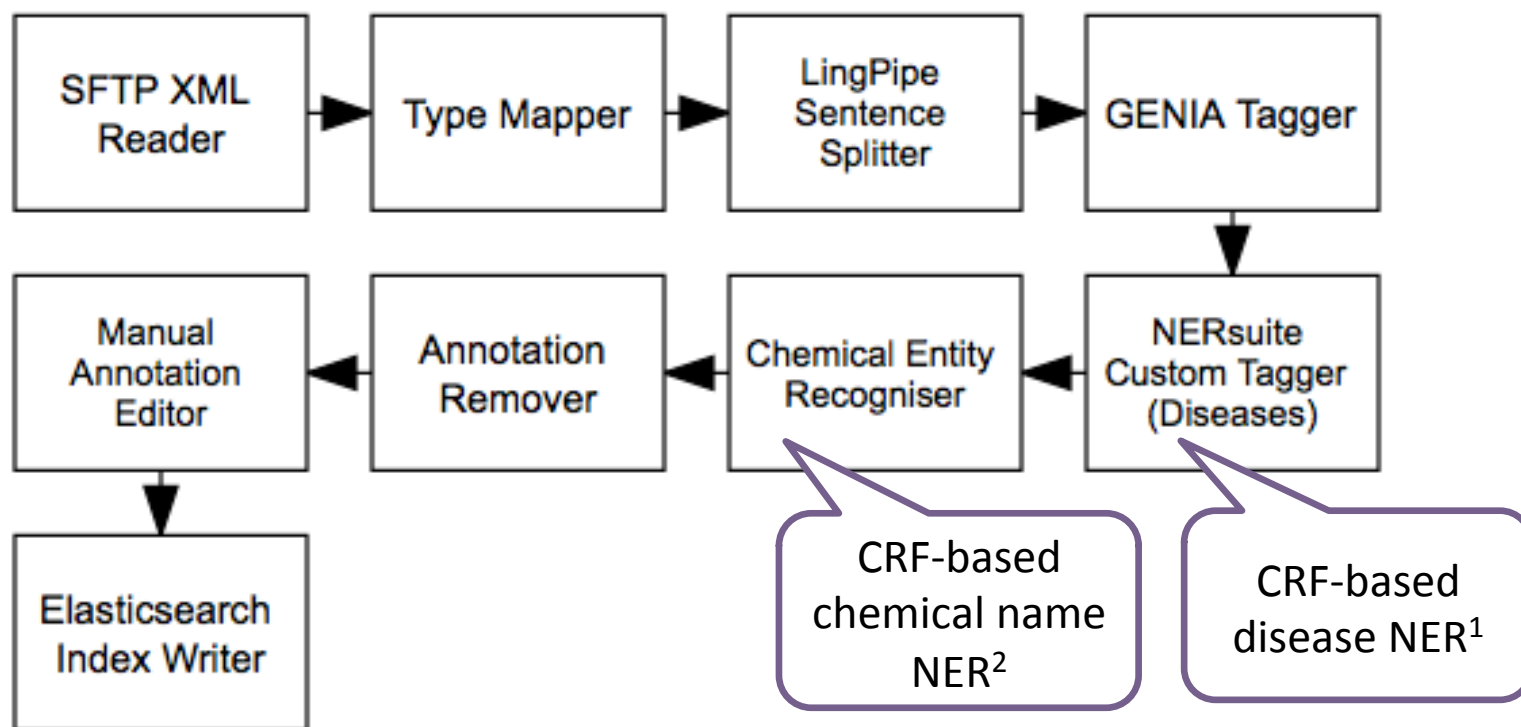
☒ **Quality**

☐ **Title**

☐ **TextBody**

Taxon
Location
Habitat
Person
Temporal expression

Generating semantic metadata with NER workflows: medical archives



¹NCBI Corpus: <http://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/>

²ChER: <https://jcheminf.springeropen.com/articles/10.1186/1758-2946-7-S1-S6>

Generating semantic metadata with NER workflows: medical archives

46



6234048.xml
6234049.xml
6234050.xml
6234965.xml
6235887.xml
6235888.xml
6235889.xml
6235890.xml
6236866.xml
6236867.xml
6236868.xml
6236869.xml
6236870.xml
6236871.xml
6237711.xml
6237712.xml
6237713.xml
6238645.xml
6238646.xml

Create

British medical journal (Clinical research ed.)
Br Med J (Clin Res Ed)

Spectinomycin as initial treatment for gonorrhoea.

1032-4

The prevalence of penicillinase producing *Neisseria gonorrhoeae* at this hospital increased exponentially from less than 0.5% in 1978 to 6.5% of all isolates in 1982. In January 1983 first line treatment for uncomplicated heterosexual anogenital gonorrhoea was therefore changed from ampicillin and probenecid to spectinomycin. This subsequently cured 95% of cases seen at the Praed Street Clinic. Although there was an initial fall in the monthly isolation rate of penicillinase producing *N gonorrhoeae* after the introduction of spectinomycin, this was not maintained. The exponential increase in the prevalence of the strain did slow in 1983, rising to only 8.7%. This, however, may have reflected a general decline in the rate of increase of penicillinase producing *N gonorrhoeae* throughout Britain. The failure to influence the prevalence of penicillinase producing *N gonorrhoeae* to any great degree may have been due in part to spectinomycin resistance in both penicillinase producing and non-penicillinase-producing

Finish Editing

AnnotationsLabels

Show ☒ Labels in document ☐ All labels

☒ Chemical

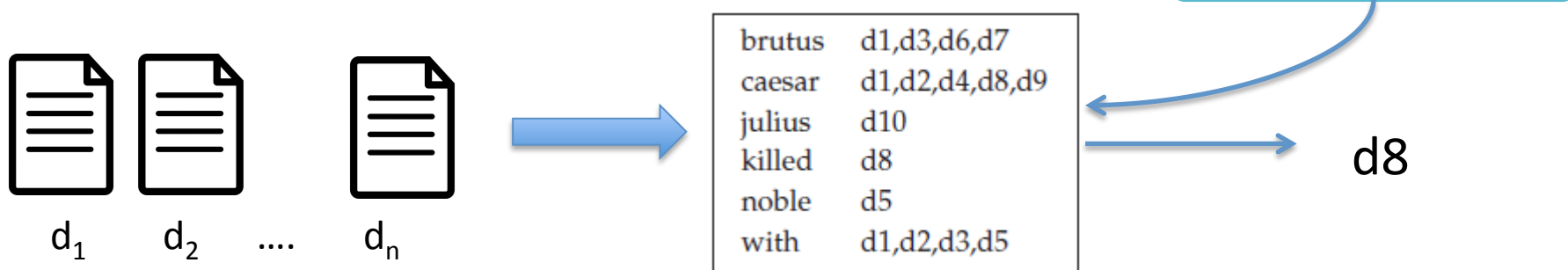
☐ TextBody

☒ Disease

Questions so far?

Introduction to Search Indices

- A **search engine** is an information retrieval system designed to help **find information stored** on a computer system
- Intuitively (and simplistically):



We will focus on Elasticsearch for this tutorial!

An overview of Elasticsearch¹

- Elasticsearch is an open-source distributed search (full-text or structured) and analytics engine:
 - timestamp or exact values,
 - full-text search, handle synonyms, score documents by relevance
 - Analytics and aggregations from the same data in real time
- Notable examples:
 - Wikipedia (full-text search, highlighted snippets, and search-as-you-type and did-you-mean suggestions)
 - The Guardian (visitor logs with social-network data to provide analytics)
 - Stack Overflow (full-text search with geolocation queries and more-like-this in Q&A)
 - GitHub (query 130 billion lines of code)
- Elasticsearch can run on your laptop, or scale out to hundreds of servers and petabytes of data

¹Much of the following content was extracted from the Elasticsearch documentation

An overview of Elasticsearch (cont)

- Built on top of Apache Lucene, a full-text search-engine library
- Lucene is arguably the most advanced, high-performance, and fully featured search engine
- Why not using Lucene then?
 - Complexity, requires a deep understanding of IR concepts and its inner workings
 - Need to work in Java and to integrate Lucene directly with your application
 - Elasticsearch packages up all this functionality into a standalone server that your application can talk to via (a RESTful) API
 - “Works right out of the box”; sensible defaults and hides complicated search theory, while still fully configurable and flexible

An overview of Elasticsearch

- Isn't Solr doing the same?
 - Which one is better depends on the application
 - Elasticsearch was born in the age of REST APIs, so it's more aligned with web 2.0 applications
 - In our case the nested document structure made Elasticsearch a clear winner
 - <http://solr-vs-elasticsearch.com>

How to install Elasticsearch

- It's quite straightforward:
 - <https://www.elastic.co/guide/en/elasticsearch/guide/current/running-elasticsearch.html>
- For development and interactive querying the recommended software is **Sense**
 - Available as a Chrome extension too
 - Send JSON data over HTTP
 - Friendly syntax for the curl command

How to communicate with Elasticsearch?

- Java API
 - Used within the Argo component
- RESTful API
 - Used for the examples here
- We will follow the ‘learn from example’ philosophy in this tutorial
 - Only emphasising important aspects of the query syntax

Elasticsearch key concepts

- Document oriented
 - Similar to the NoSQL concept of *document*
 - Intuitively, a document is analogous to an object in OO-programming
 - Why? No need to squeeze or flatten your object into a table (usually one field per column) losing its richness
- JSON
 - Serialisation format for documents

```
{
  "email":      "john@smith.com",
  "first_name": "John",
  "last_name":  "Smith",
  "info": {
    "bio":      "Eco-warrior and defender of the weak",
    "age":      25,
    "interests": [ "dolphins", "whales" ]
  },
  "join_date": "2014/05/01"
}
```

Elasticsearch key concepts

- Glossary:
 - Index:
 - analogous to a database in SQL and NoSQL
 - can contains multiple **types**
 - **Type**:
 - analogous to a table (SQL) or collection (MongoDB)
 - can contain multiple **documents**
 - **Document**:
 - analogous to a row (SQL)
 - can contain multiple **fields**
 - **Field**:
 - Analogous to a column (SQL)
 - Each field is associated with a field type: 'string', 'date', 'integer'
- Index is an overloaded word
 - as a *noun*, as a *verb* and *inverted index*

Querying Elasticsearch

- We already ran Argo workflows, which inserted data in Elasticsearch
- Let's have a look at the existing indices...
- Let's search for all documents in an index...
 - Format of the response
 - Pagination

Querying Elasticsearch

- Let's refine the query searching for a specific term...
- Let's search for entities...

Querying Elasticsearch using Sense

```

5
6 GET /medical_archives/_search
7
8 GET /medical_history/_search
9
10 GET /medical_history/document/_search
11 {
12   "size": 20,
13   "query": {
14     "match": {
15       "text": "AIDS"
16     }
17   }
18 }

```

```

20 # search for annotations!
21 GET /medical_history/document/_search
22 {
23   "query": {
24     "nested": {
25       "path": "metadata",
26       "query": {
27         "bool": {
28           "must": [
29             {
30               "match": {
31                 "metadata.obj": "Disease"
32               }
33             },
34             {
35               "match": {

```

```

946     "obj": "Disease",
947     "word_form": "motor neurone disease"
948   }
949   ]
950   }
951   },
952   {
953     "_index": "medical_history",
954     "_type": "document",
955     "_id": "AVcG9YXp2vLZYWctVilx",
956     "_score": 0.8829731,
957     "_source": {
958       "text": "\n
A patient was admitted to hospital with an apparent psychiatric
disturbance. When she became stuporous the cerebrospinal fluid was cultured but proved sterile. The latex test
showed that serum was positive for cryptococcal antigens, and cryptococcal meningoencephalitis was diagnosed.
Amphotericin B was given but when she developed a toxic reaction it was replaced by flucytosine. She responded
well to flucytosine alone and no side effects appeared on continued treatment. Cryptococcal meningitis may
present as a psychiatric disturbance, and serological tests are invaluable aids to diagnosis.\n
",
959     "metadata": [
960       {
961         "obj": "Source",
962         "word_form": "1095135.xml"
963       },
964       {
965         "span": {
966           "begin": 69,
967           "end": 92
968         },
969         "obj": "Disease",
970         "word_form": "psychiatric disturbance"
971       },

```

Some caveats

- No need to define a mapping (i.e. schema)
 - Elasticsearch tries to guess it (“works out of the box”)
 - But in most cases it is necessary to define it:
 - Define nested objects as such (e.g. ‘metadata’)
 - Define fields that do not need text processing (e.g. metadata fields)
- Let’s have a look at our current mappings...

Much more...

- Aggregation (faceting)
- Horizontal scalability (sharding)
- Sorting / relevance
- Word proximity, partial matching, fuzzy matching, and language awareness
- Geolocation and geohashes

Questions so far?

- <http://nactem.ac.uk/hom>
- Archives
 - British Medical Journal articles (380,000)
 - London Medical Office of Health reports (5,000)

Search

My Documents

Full Access ?

Search Query

You are currently viewing the entire set of documents. To find specific documents please refine your search.

Start refining search

Search Results ?

Showing page 1 of 385409 results

- [icon]

Cardiomyopathy associated with Wegener's granulomatosis
Cardiomyopathy associated with Wegener's granulomatosis Andrew To Janak De Zoysa Jonathan P Christiansen Jonathan.Christiansen@WaitemataDHB.govt.nz KEYWORDS: A 35-year-old man with no history of cardiovascular disease presented with severe biventricular heart failure, after a 6-week history ...
bmj_3106061

+ [icon]

Recurrent wheezing ... is it only asthma?
Recurrent wheezing ... is it only asthma? Luis Vaz Rodrigues Cristina Lopes a Castel-Branco 1 Centro Hospitalar de Coimbra, Pulmonology, Quinta dos Vales, S Martinho do Bispo, Coimbra 3041-801, Portugal 2 Hospital de São João, EPE, Immuno-allergology Division, Alameda Prof Hernâni Monteiro, Porto...

Searching for “cold” based on keywords

Search My Documents

Full Access

Search Query

Term ×
cold

New Search

Refine Search

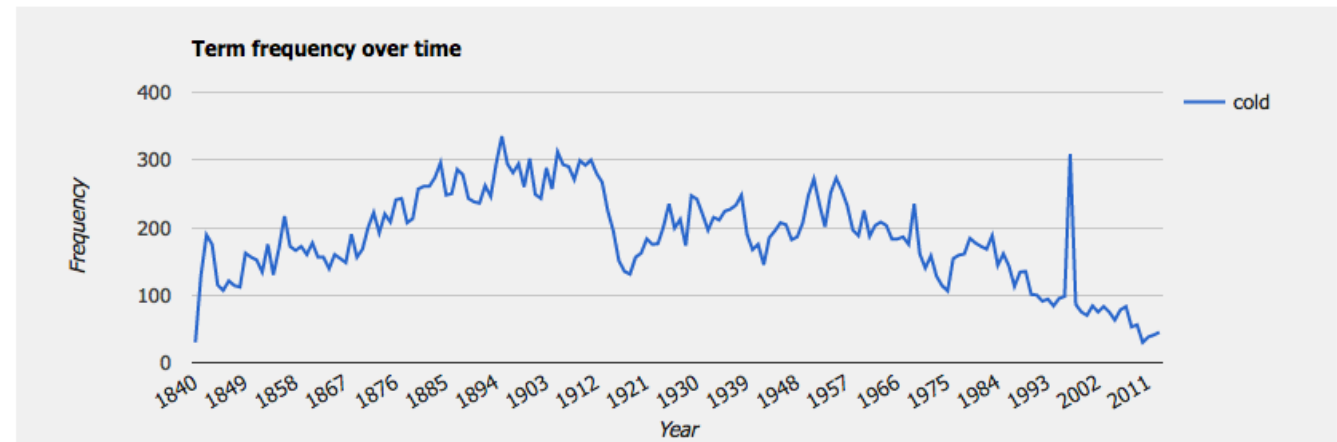
Related Terms

boil radiator cool supply
evaporation
fluoridation closet drinking bucket
wineglass
brackish hot soap wash unfiltered
filter
moorland drink
cooler
warm

Search Results ?

Showing page 1 of 30885 results

Term Frequencies



The Common Cold

The Common Cold SIR,-The following facts may be of interest to either sufferers from or investigators of the common cold. During the course of twenty-five years' practice of psycho-analysis for the treatment of psychoneurosis I have observed that in them: (1) A cold invariably occurred in a part...

bmj_2285094

Searching for “cold” based on keywords



The Common Cold

The Common Cold SIR,-The following facts may be of interest to either sufferers from or investigators of the common cold. During the course of twenty-five years' practice of psycho-analysis for the treatment of psychoneurosis I have observed that in them: (1) A cold invariably occurred in a part...

bmj_2285094

“Cold” as a medical condition



POINTS FROM LETTERS: The Common Cold

POINTS FROM LETTERS The Common Cold Dr. E. WRIGLEY BRAITHWAITE (Leeds) writes: In 1943 (Journal, 2, 433) you published a letter from me in which I made a number of statements about the pathology of the common cold. . . . It only develops when the individual is in a particular emotional state, wh...

bmj_2039150



Why “A Cold”?

Why “A Cold”? SIR,-I have read Mr. James Crooks's article on nasal sin sinusitis in childhood (April 30, p. 935). In it the expression “a cold,” without any definition whatever, is repeatedly used. I suggest that the terms “a cold,” “catching cold,” and “a chill” should be dropped from me...

bmj_2086387

“Cold” to describe temperature



Cold Drink and Syncope

770 BRITISH MEDICAL JOURNAL 23 September 1972 Cold Drink and Syncope St.-Flight Lieutenant D. J. Rainford's letter (19 August, p. 475) reminds me of a verse I learnt at my mother's knee (she came from a medical family). ...

bmj_1788630



Feeling the Cold

Feeling the Cold SIR,-Your annotation on “Feeling the Cold” (February 21, p. 445) tells us that “cold and warmth are separately appreciated. Different points on the skin respond either to cold or to

Searching for “cold” as a disease based on semantic metadata

65

[Search](#) [My Documents](#) Full Access ?

Search Query

Condition ×

cold

[New Search](#) [Refine Search](#)

Search Results ?

Showing page 1 of 3825 results

+ 📄

Diffuse pleural thickening and diaphragmatic paralysis causing combined intrathoracic and extrathora...

Diffuse pleural thickening and diaphragmatic paralysis causing combined intrathoracic and extrathoracic pulmonary restriction John J Dixon Grania J Price F Runa Ali 1 St. Helier Hospital, Nephrology, Wrythe Lane, Carshalton, London, SM5 1AA, UK 2 Guy's Hospital, Anaesthetics, Great Maze Pond, L...

bmj_3029080

+ 📄

Effect of point of care testing for C reactive protein and training in communication skills on antib...

Effect of point of care testing for C reactive protein and training in communication skills on antibiotic use in lower respiratory tract infections: cluster randomised trial Jochen W L Cals - general practitioner trainee and researcher Christopher C Butler - professor of primary care medicine Rog...

bmj_2677640

+ 📄

Monitoring the emergence of community transmission of influenza A/H1N1 2009 in England: a cross sect...

Monitoring the emergence of community transmission of influenza A/H1N1 2009 in England: a cross sectional opportunistic survey of self sampled telephone callers to NHS Direct Alex J Elliot - project lead Cassandra Powers - scientist Alicia Thornton - scientist Chinelo Obi - research assistant C...

bmj_2733951

+ 📄

Effect of using an interactive booklet about childhood respiratory tract infections in primary care ...

Effect of using an interactive booklet about childhood respiratory tract infections in primary care

Applications: BHL Query Expansion

- <http://nactem10.mib.man.ac.uk/va/MiBio/Search/queryExpansion.html?prot=thumb>

Rock pigeon

Search results

The variation of animals and plants under domestication / by Charles Darwin ; authorized edition...

Orange Judd & Co., - 1868.

... **rock-pigeon**. In fantails, which have their tails so largely developed, there are either eight or nine ... , and apparently in one case ten, and they are a little longer than in the **rock-pigeon**, and their shape ... little more equally rounded on both sides than in the **rock-pigeon**. The ischium is also frequently rather ...

The variation of animals and plants under domestication /

D. Appleton and company, - 1883.

... of the **rock-pigeon**, calculated (with a few specified exceptions) by the standard of the length of the ... , being a Short-faced Tumbler, is much smaller than the **rock-pigeon**, and would naturally have shorter feet ... of the **rock-pigeon**, relatively to the size of the body in these two birds, as measured from the base ...

On the origin of species by means of natural selection, or, The preservation of favoured races in...

John Murray ..., - 1859.

... consideration. The **rock-pigeon** is of a slaty-blue, and has a white rump (the Indian subspecies, C. intermedia of ... any wild **rock-pigeon** ! We can understand

Query

Sort by: ☒ Relatedness ☐ Frequency



Frequency

Rock pigeon

You might also be interested in...



Relatedness

Frequency

Rock dove



Relatedness

Frequency

Common pigeon

Searching for “Aquila chrysaetos”

Aquila chrysaetos

Search results

Natural history of the animal kingdom for the use of young people : in three parts, comprising I....

E. & J.B. Young and Co., - 1889.

... [Accipitres. Eagles and Kite. a) Golden Eagle. **Aquila chrysaetos**. b) White-tailed Eagle ...

Osteology of birds, by R.W. Shufeldt.

University of the State of New York, - 1909.

... [31 Right lateral view of the trunk skeleton of the Golden eagle (**Aquila chrysaetos**). Reduced one ...

Annotated bibliographies on selected bird species inhabiting the California desert /

s.n., - c 1979]

... [» **Aquila chrysaetos** (con't.) low, it is at the apex of a food chain, and its large body size 1978. Bibliography on the golden eagle (**Aquila chrysaetos**) Revised ed. Published by Raptor Research ... Foundation. Snow, C. 1973. Golden eagle **Aquila chrysaetos** . Habitat management series for endangered species ...

Check-list of birds of the world.

Harvard University Press, - 1931-1987.

... , 1924, p. 59-60. Swann, Bull. Brit. Orn. Cl., 45, 1925, p. 64-73. **Aquila chrysaetos** chrysaetos (Linne ... , southern Russia, Caucasus and Asia Minor.



152 entries found
Page 1 of 8



Query

Sort by: ☒ Relatedness ☐ Frequency



Frequency

Aquila chrysaetos

You might also be interested in...



Relatedness

Frequency

Golden eagle



Relatedness

Frequency

Black eagle



Searching for “Aquila chrysaetos”: expanding with “Golden eagle”

Aquila chrysaetos

Expanded query:

"aquila chrysaetos" OR "Golden eagle"

Search results

Natural history of the animal kingdom for the use of young people : in three parts, comprising I....

E. & J.B. Young and Co., - 1889.

... [Accipitres. Eagles and Kite. aj **Golden Eagle. Aquila chrysaetos.** b) White-tailed Eagle ...

Osteology of birds, by R.W. Shufeldt.

University of the State of New York, - 1909.

... [31 Right lateral view of the trunk skeleton of the **Golden eagle (Aquila chrysaetos)**. Reduced one ...

Annotated bibliographies on selected bird species inhabiting the California desert /

s.n., - c 1979]

... [**Aquila chrysaetos** (con't.) Sumner, E.L., Jr. 1929c. **Golden eagle** in Death Valley. Condor 31:127 Taylor, H.R. 1890. Nesting habits of the **golden eagle**. Zoe 1:42-44. Thelander, C.G. 1974. Nesting ... territory utilization by golden eagles (**Aquila chrysaetos**) in California during 1974. Wildlife Management ...

The birds of Albany County; a catalogue of the species recorded in this vicinity with notes on th...

[Press of Bradow printing co.] - 1907.

... . American Hawk Owl. — *Burnia alula capareoch.* **Golden Eagle.** — **Aquila**



152 entries found + 371 entries added

Page 1 of 27



Query

Sort by: ☒ Relatedness ☐ Frequency



Frequency

Aquila chrysaetos

You might also be interested in...



Relatedness

Frequency

Golden eagle



Relatedness

Frequency

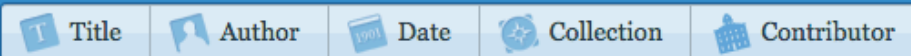
Black eagle



Searching for “Aquila chrysaetos” in BHL



Browse by:



Search



[ADVANCED SEARCH](#)

Results for "Aquila chrysaetos"

Books/Journals (0)

Articles/Chapters/Treatments (4)

Authors (0)

Subjects (1)

Scientific Names (12)

Sort By: Relevance Title Author Year

[Download results](#)

[GOLDEN EAGLE AQUILA-CHRYSAETOS BREEDING IN OMAN EASTERN ARABIA](#)

By: Gallagher, M D ;Brown, M R

Type: Article

In: Bulletin of The British Ornithologists' Club

Volume: 102

Date: 1982

[View Article](#)

[On the Breeding of the Golden Eagle \(Aquila chrysaetos\) in North-western India](#)

By: Unwin, W H

Type: Article

In: Proceedings of The Zoological Society of London

Volume: 1874

Date: 1874

[View Article](#)

SUPPORT

Help Support BHL

BHL's existence depends on the support of its patrons. Help us keep this free resource alive!

[Donate Now](#)

Featured Content
BHL at 10

Conclusions

- Discussed challenges in information discovery and search
- Reviewed methods for NER
- Presented the Argo text mining workbench
- Extracted named entities which are then indexed to facilitate semantic searches
- Presented fundamentals of Elasticsearch: key concepts, search, mappings

Conclusions

- Illustrated some applications:
 - Disambiguation in the History of Medicine system
 - Improving recall in BHL
- Please get in touch with us if you're interested in applying Argo to your digital libraries!
 - riza.batista@manchester.ac.uk
 - axel.soto@manchester.ac.uk