BEL.bio Semantic Terminology Services

William Hayes, PhD
whayes@biodata.com

Mission

- Provide Terminology services for BEL

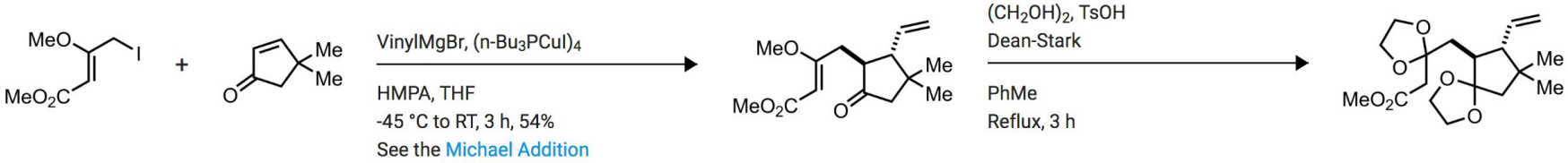
- Target audience:
  - General Biologists
  - Computational Biologists
  - Toxicologists

Goals

- Powerful terminology service
  - Very fast, scalable term search
  - Completion support for terms
- Management of terminology datasets
- Flexible approach to collect and normalize terminologies from sources
- Open-source friendly approach
  - Easier to get started, not too complex an ecosystem, modular
- Easy to update, limit any downtime to update

What is BEL?

- Chemists have the Chemical Reaction Language



Partial chemical synthesis pathway:
https://www.synarchive.com/syn/128

- Biologists now have Biological Expression Language (BEL)
  - Open standard for communication and knowledge-storage
  - Whiteboard and Computer friendly

What is BEL, cont'd?

- BEL – Biological Expression Language
- Pubmed Abstract: "… Northern blot analysis documented that two transcription factor genes chosen for further study, c-myc promoter-binding protein (MBP-1) [official symbol: ENO1] and X-box binding protein 1 (XBP-1), were up-regulated in U266 cells about 3-fold relative to the cell cycle-dependent beta-actin gene 12 h after IL-6 treatment …"
- BEL Assertions
  - *p(HGNC:IL6) increases r(HGNC:ENO1)*
  - *p(HGNC:IL6) increases r(HGNC:XBP1)*
- *Annotations*
  - *Species: Human*
  - *CellLine: U266*

What is BEL.bio?

- BEL.bio
  - http://bel.bio
  - https://github.com/belbio
- Open source project to provide BEL support and tools
  - BEL validation
  - BEL and term completion
  - BEL language parsing and manipulation
- Terminologies are a critical part of BEL
  - *Many different entity or concept types*
  - *Canonicalization*
  - *Equivalencing*
  - *Hierarchical terminologies*
  - *Orthologization support*

BEL Terminology processing history

- Version 1: Started with custom formats for terminologies (OpenBEL *.belns, *.beleq), required building equivalencing, etc into a monolithic dataset for efficient equivalencing – RDBMS (Java-based)

- Version 2: Used SKOS, MongoDB for completions, Semantic Web Triplestore (Java, Jruby, and Ruby)

- Version 3: JSON format, Elasticsearch, ArangoDB (Python)

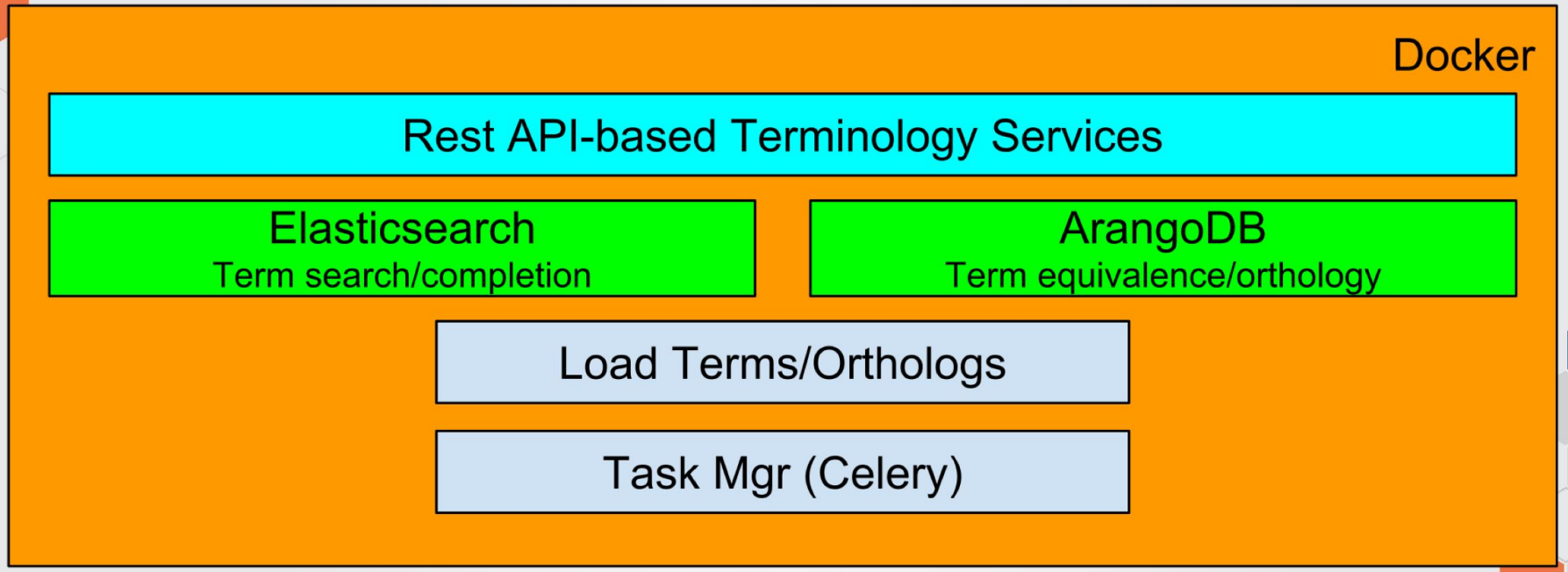
BEL Terminology processing history

- Version 1: Started with custom formats for terminologies (OpenBEL *.belns, *.beleq), required building equivalencing, etc into a monolithic dataset for efficient equivalencing – RDBMS (Java-based)

- Version 2: Used SKOS, MongoDB for completions, Semantic Web Triplestore (Java, Jruby, and Ruby)

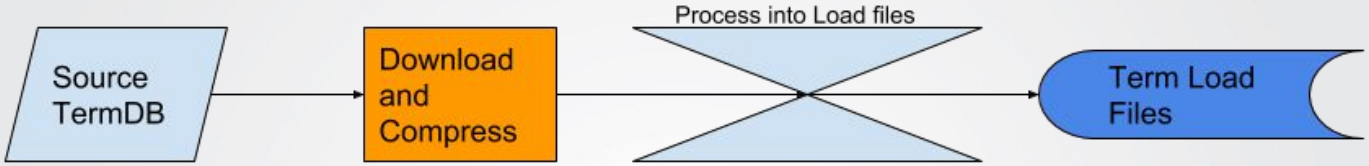
- Version 3: JSON format, Elasticsearch, ArangoDB (Python)

Architecture

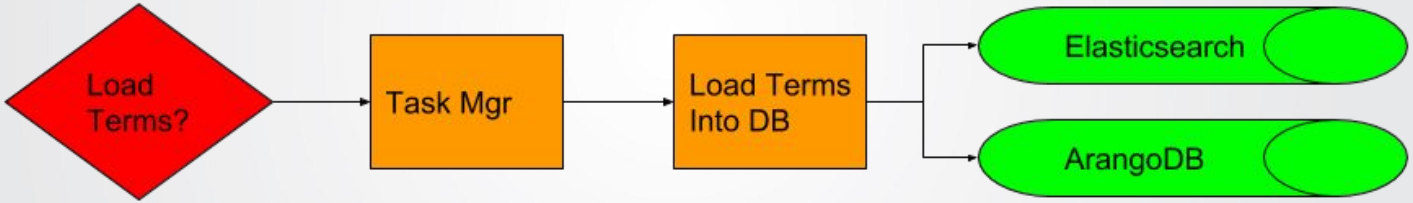


Terminology Workflow

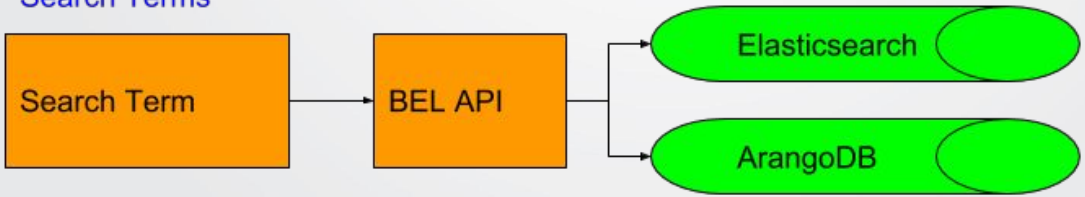
Generate Term Load Files



Load Terms into BEL API



Search Terms



- Elasticsearch – premier search-oriented docstore
  - Provides term search support filtered by any term metadata (e.g. species)
  - Provides term completion support using XXX

- ArangoDB – excellent distributed key-value, doc, graphdb docstore
  - Supports Equivalencing graph queries
  - Supports Orthologization graph queries

Loading Terminologies

- Reduce downtime when loading terminologies
- Load terms into new Elasticsearch index (term_<namespace>_isodate)
- Remove old Elasticsearch index and alias new index to ‘terms’

- ArangoDB is harder
- Load ArangoDB equivalence/orthology documents as nodes and edges
- Add source and update datetime to nodes and edges
- Remove all old documents with same source and older update datetime

- Admin User starts new load by sending a POST to <bel_api_url>/tasks/resources with URI of Terminology file
- Terminology file gets added to Task queuing service (Celery)

JSONLines format

- JSONLines Format: http://jsonlines.org/
- One line per JSON record
- Compresses well (e.g. HGNC 18Mb to 1.9Mb compressed)

- {"metadata": {"name": "HGNC", "type": "namespace", "namespace": "HGNC", "description": "Human Gene Nomenclature Committee", "version": "2018-04-20T13:34:22", "src_url": "http://www.genenames.org", "url_template": "http://www.genenames.org/cgi-bin/gene_symbol_report?hgnc_id=<src_id>"}}
- {"term": {"namespace": "HGNC", "namespace_value": "A1BG", "src_id": "5", "id": "HGNC:A1BG", "alt_ids": ["HGNC:5"], "label": "A1BG", "name": "alpha-1-B glycoprotein", "species_id": "TAX:9606", "species_label": "human", "description": "", "entity_types": ["Gene", "RNA", "Protein"], "equivalences": ["SP:P04217", "EG:1"], "synonyms": [], "children": [], "obsolete_ids": []}}

- This format is memory efficient when processing large terminologies. Metadata must be first line in JSONLines terminology file.

Terminology JSON Structure

- Metadata section (name, src, namespace, version, ...)
- Term section (examples from HGNC:AKT1)
  - Namespace (HGNC)
  - Namespace value (AKT1)
  - ID (HGNC:AKT1)
  - Source ID (AKT1)
  - Alternate IDs (HGNC:391)
  - Label (AKT1)
  - Name (AKT serine/threonine kinase 1)
  - Species ID TAX:9606
  - Species Label Human
  - Description
  - Entity types (Gene, RNA, Protein)
  - Annotation types
  - Equivalences (SP:P31749, EG:207)
  - Synonyms (RAC, PKB, PRKBA, AKT, v-akt murine thymoma viral oncogene homolog 1)
  - Children
  - Obsolete IDs

https://github.com/belbio/schemas/tree/master/schemas/{terminology*}

Term JSON example


```
"term": {
  "namespace": "HGNC",
  "namespace_value": "AKT1",
  "src_id": "391",
  "id": "HGNC:AKT1",
  "alt_ids": [
    "HGNC:391"
  ],
  "label": "AKT1",
  "name": "AKT serine/threonine kinase 1",
  "species_id": "TAX:9606",
  "species_label": "human",
  "description": "",
  "entity_types": [
    "Gene",
    "RNA",
    "Protein"
  ],
```


<<<continued on right>>>


```
  "equivalences": [
    "SP:P31749",
    "EG:207"
  ],
  "synonyms": [
    "RAC",
    "PKB",
    "PRKBA",
    "AKT",
    "v-akt murine thymoma viral oncogene homolog 1"
  ],
  "children": [],
  "obsolete_ids": []
}
```


https://github.com/belbio/schemas/tree/master/schemas/{terminology*|ortholo
gy*}

Equivalence Results

- ArangoDB Query:
  - FOR vertex, edge IN 1..10
    - ANY 'equivalence_nodes/EG:207' equivalence_edges
    - RETURN DISTINCT {term_id: vertex._key, namespace: vertex.namespace}
- Using a graph database means that we can add at will and don't have to build a global equivalences file from all terminologies

term_id	namespace
EG:207	EG
SP:P31749	SP
HGNC:AKT1	HGNC

Term Completion Examples – BioDati Studio

NANOPUB EDITOR

CITATION

CONTENT (01)

METADATA (00)

PUBLISH | MEDIUM | HUMAN | BEL: 2.0.0

Assertions (00)

SUBJECT
p(akt

Relation

Object

CREATE

HGNC:AKT1 (AKT1) *AKT*

SP:Q9Y243 (AKT3) Protein kinase *Akt*-3

EG:207 (AKT1) *AKT*

SP:P31751 (AKT2) Protein kinase *Akt*-2

HGNC:AKT2 (AKT2) *AKT* serine/threonine kinase 2

HGNC:AKTIP (AKTIP) *AKT* interacting protein

HGNC:AKT1S1 (AKT1S1) *AKT1* substrate 1

HGNC:AKT3 (AKT3) *AKT* serine/threonine kinase 3

Description

Denotes the abundance of a protein

Function Summary

proteinAbundance(NSArg, loc|frag()?, var|pmod()*)

Argument Help

1. Namespace argument of following type(s): Protein
2. Zero or one of each function(s): location, fragment
3. Zero or more of each function(s): variant, proteinModification

Term Completion Examples – BioDati Studio

NANOPUB EDITOR

CITATION

CONTENT (01)

METADATA (00)

PUBLISH | MEDIUM | HUMAN | BEL: 2.0.0

Assertions (00)

SUBJECT
path(mu

Relation

Object

CREATE

MESH:"Heavy Chain Disease" (Heavy Chain Disease) *mu* Chain Disease

MESH:Mutism (Mutism) Elective *Mutism*

MESH:"Multiple Chronic Conditions" (Multiple Chronic Conditions) *Multiple* Chronic Conditions

MESH:"Lipomatosis, Multiple Symmetrical" (Lipomatosis, Multiple Symmetrical) Lipomatosis, *Multiple* Symmetrical

MESH:"Multicystic Dysplastic Kidney" (Multicystic Dysplastic

Description

Denotes a disease or pathology process

Function Summary

pathology(NSArg)

Argument Help

1. Namespace argument of following type(s): Pathology

Term Completion Examples – BioDati Studio

Annotations (01)

TYPE: Anatomy

ID: lung

Label

CANCEL | CREATE

Edit | Delete

- [ ] **Type**
- [ ] Species

ID: MESH:Lung LABEL: Lung MATCHED: *Lung*

ID: UBERON:lung LABEL: lung MATCHED: *lung*

ID: MESH:"Extravascular Lung Water" LABEL: Extravascular Lung Water MATCHED: MESH:"Extravascular *Lung* Water"

ID: UBERON:"lung hilus" LABEL: lung hilus MATCHED: UBERON:"*lung* hilus"

ID: UBERON:"lung elastic tissue" LABEL: lung elastic tissue MATCHED: UBERON:"*lung* elastic tissue"

ID: UBERON:"right lung lobe" LABEL: right lung lobe MATCHED: UBERON:"right *lung* lobe"

ID: UBERON:"lung bud" LABEL: lung bud MATCHED: UBERON:"*lung* bud"

ID: UBERON:"lung field" LABEL: lung field MATCHED: UBERON:"*lung* field"

ID: UBERON:"lung parenchyma" LABEL: lung parenchyma MATCHED: UBERON:"*lung* parenchyma"

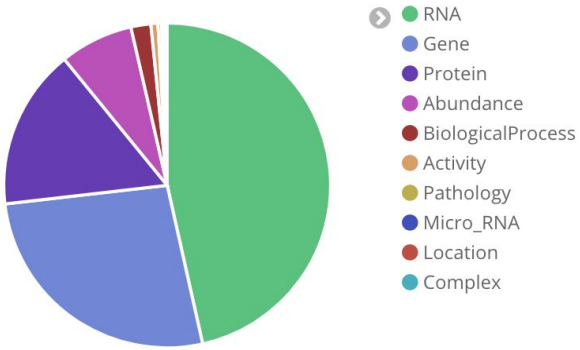
ID: UBERON:"right lung" LABEL: right lung MATCHED: UBERON:"right *lung*"

Kibana Namespace Dashboard – Human, Mouse, Rat, Zebrafish

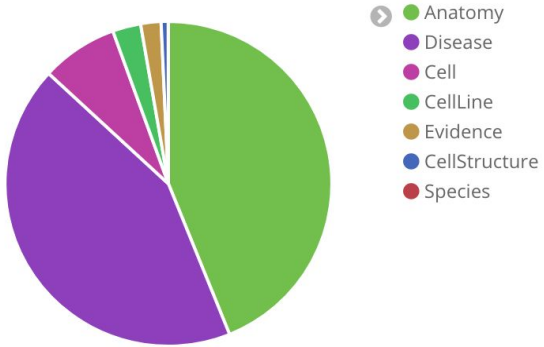
Namespace Counts

namespace: Descending	Count
AFFX	327,392
EG	219,122
CHEBI	106,644
MGI	57,532
SP	48,352
RGD	44,972

Entity types



Annotation types



Namespace Statistics – All species

Namespaces	Count
EG	20,750,186
TAX	1,736,298
SP	557,012
AFFX	327,392
CHEBI	106,644
MGI	57,532
RGD	44,972
GO	44,922
HGNC	41,315
ZFIN	23,388
MESH	19,223
UBERON	13,232
DO	8,699
CL	2,194
EFO	937

More information

- BioDati – http://biodata.com
- BEL.bio – http://bel.bio
- Github - https://github.com/belbio
- What is BEL? - https://medium.com/biodati/what-is-bel-8df1a549760f
- BEL Namespace Completion - https://medium.com/biodati/bel-namespace-completion-79ce5501af81 (overview of the approach used)

BioDati, Inc Applications and Services

- Applications
  - BioDati Studio – open-standard based (BEL) biological network visualization tool and editor
    - Manage biological knowledge (Nanopubs)
    - Build biological networks from imported or internally-curated BEL
  - NetworkStore
  - NanopubStore
- Services:
  - BEL resource and tool support (Namespaces, database conversion, BEL format conversions)
  - BEL training
  - BEL.bio open source development
  - Terminology platform and content services
  - Consulting

Acknowledgements

- Natalie Catlett, PatientsLikeMe
- Anselmo DiFabio, BioDati
- David Chen
- Tony Bargnesi
- Nick Bargnesi

email: support@biodata.com | online: biodata.com

Rahway, NJ and Boston, MA