BioSolr

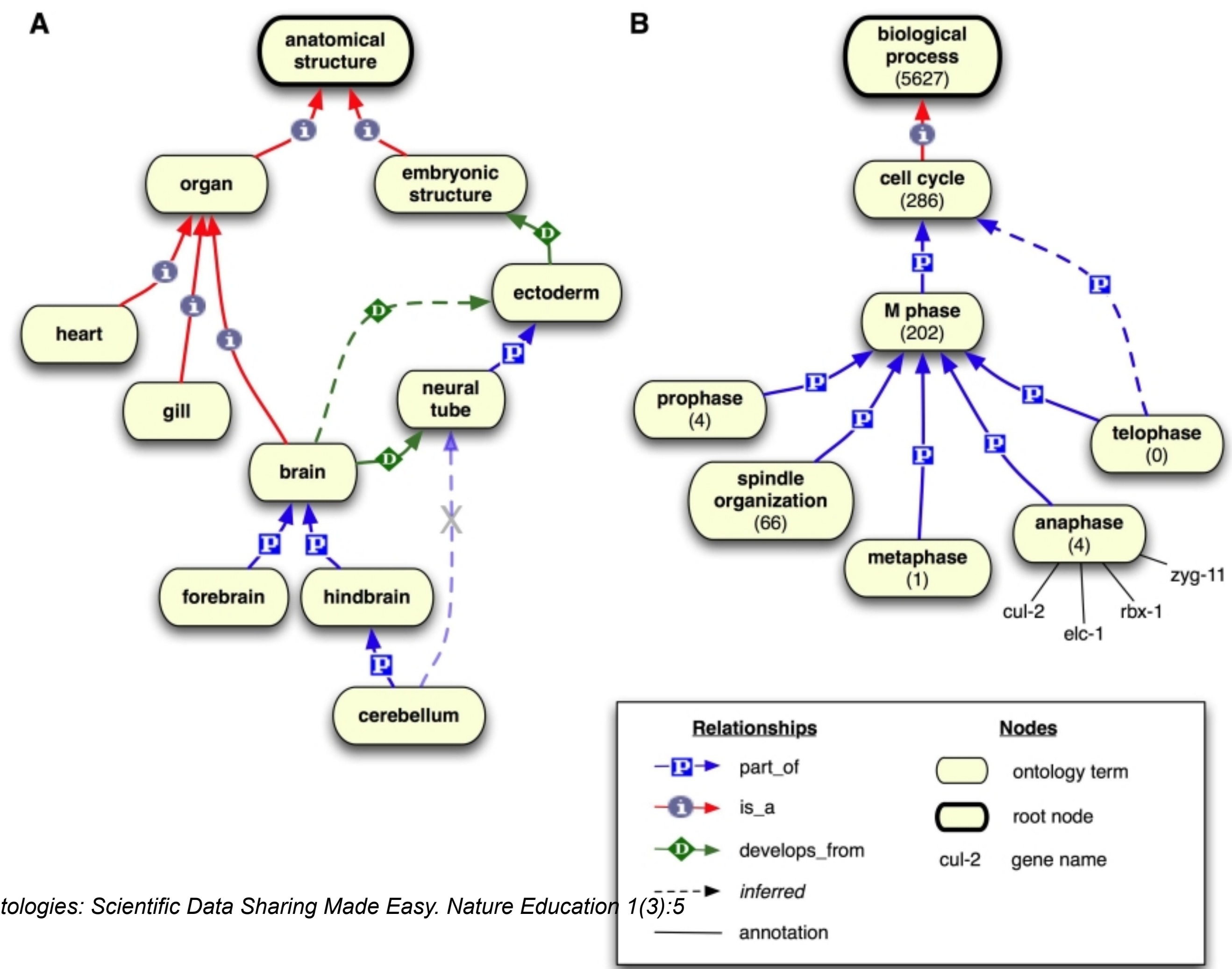*building better search for bioinformatics*

Elasticsearch Workshop Day
February 2016

Charlie Hull - Managing Director
Matt Pearce – Senior Developer

charlie@flax.co.uk
www.flax.co.uk/blog
+44 (0) 8700 118334
Twitter: @FlaxSearch

BBSRC
20 Years of Pioneering
Great British Bioscience

flax
The Open Source Search Specialists

EMBL-EBI

BioSolr & SPOt – Indexing ontologies

Washington, N. & Lewis, S. (2008) Ontologies: Scientific Data Sharing Made Easy. Nature Education 1(3):5

Indexing ontologies – the problem

- You have a collection of documents annotated with ontology references.

- You want to search both the documents and the associated ontology data.
  - This may include associated nodes – "has location", "is part of", etc.

- Faceting the ontology references would be nice!
  - (especially if the facets can be presented in a tree)

Approach 2 – final result

- We have developed a Solr UpdateProcessor to do this as part of the update ch
- The user defines the ontology location (file, URL) and the field to reference.
- Aims for convention over configuration for remaining properties.
  - Field names and ontology annotation details all customisable.

- *A similar plugin for ElasticSearch has also been developed
  - we'll talk about this tomorrow!;*

ElasticSearch ontology annotator plugin

- Implemented as a mapper - a new field type.

- Adds new sub-fields to the annotation field.

- Configuration in field mapping, rather than index settings.

ElasticSearch ontology annotator plugin (2)

- Built on same ontology look-up code as the Solr version.

- (although code is currently duplicated, since we haven't published any mo

- Can use OWL/RDF files as base, or OLS API..

Differences from Solr implementation

- Configuration is (probably) simpler

- Adds sub-fields to the annotation field, rather than adding top-level fields.

- However, sub-fields not added to _source, need to be specified in query fi

- Installs using Elasticseach plugin manager, rather than by copying JARs

Issues with this implementation

- ElasticSearch is constantly evolving

- Different code requirements for most versions of ES...
  - so we're aiming at a moving target, module will require constant update

- Having to specify the new fields in your retrieval list is a pain.

- ElasticSearch community seem less supportive of plugin development tha

Get Involved!

- Check out the github page: https://github.com/flaxsearch/BioSolr

- Suggest more use cases for what we've built!

Thank you for listening – any questions?

.co.uk/blog

700 118334

@flaxsearch