Open source search for Bioinformatics

EMBL-EBI 3-4 Feb 2016

Acknowledgements

- Gautier Koscielny
- Tony Burdett
- Helen Parkinson
- Charlie Hull
- Matt Pearce
- Tom Winch

- BBSRC Flexible Interchange Programme (FLIP) grant number BB/M013146/1

Genesis of BioSolr

- Grant Ingersoll visits the Wellcome Campus in July ’13
- Around 90 people attend
- Show of hands indicates 75% using Lucene/Solr

- Still remember Charlie and Grant’s face when they saw the data-config file I had created

BioSolr

"BioSolr aims to significantly advance the state of the art with regards to indexing and querying biomedical data with freely available open source software"

One year BBSRC funded project from September 2014 (since extended to February 2016)

BioSolr team

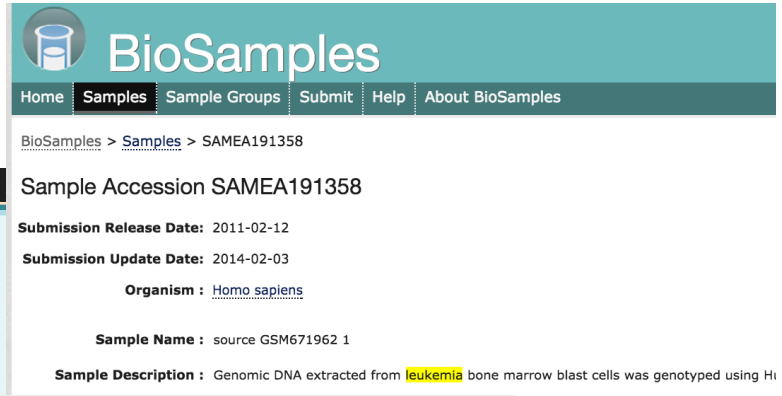
- Tom Winch
  - Working on site with Sameer Velankar & the PDBe team
  - Facet.contains, Xjoin, Federated Search
- Matt Pearce
  - Working on site with Tony Burdett & the SPOT team
  - Indexing ontologies

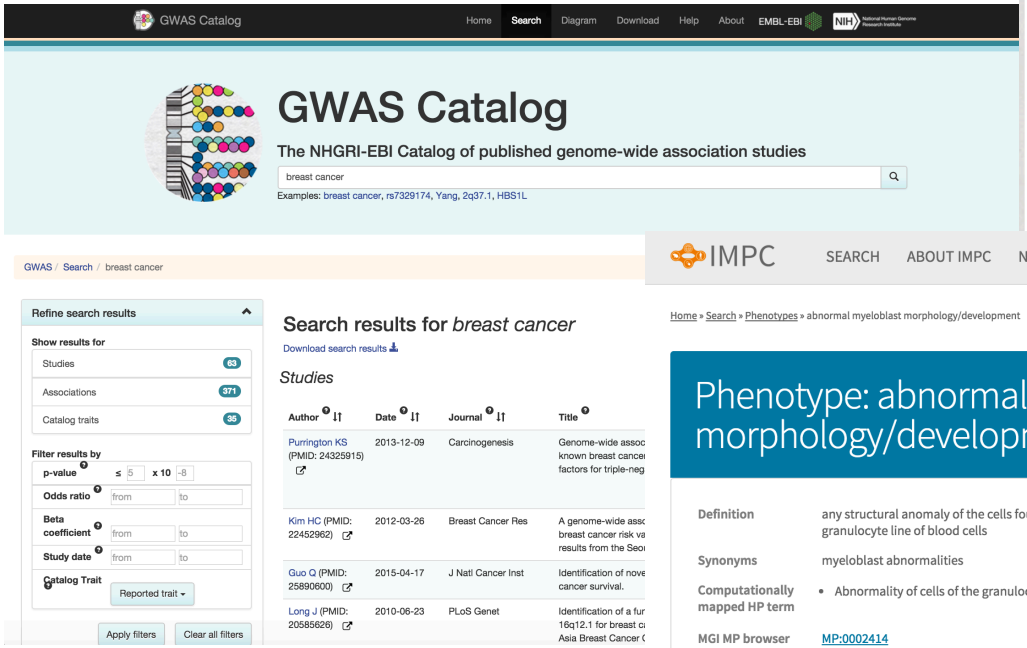
SPOT Team

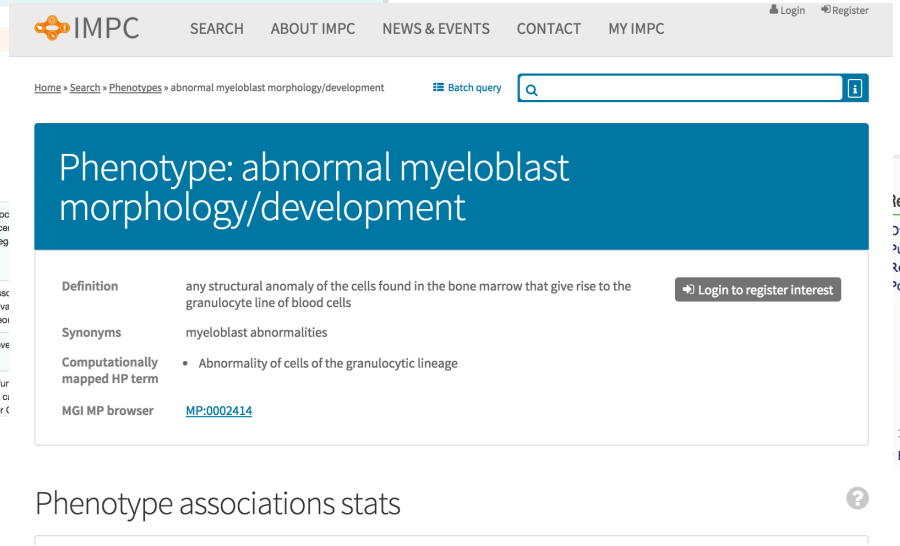
- Three themes:
  - Biosamples and Semantic Integration
  - Mouse Informatics
  - Gene Ontology Editorial Office
- This means we do lots of:
  - Ontology building (especially application ontologies – GO, EFO, CMPO)
  - Ontology annotation and mapping
  - Development of tools to support these goals (OLS, Zooma, Webulous and more)

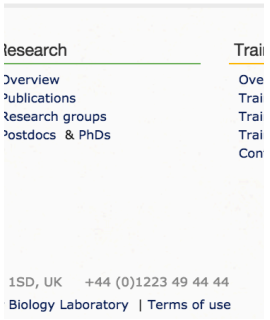
Searching

- We provide several databases that make use of semantically enriched search

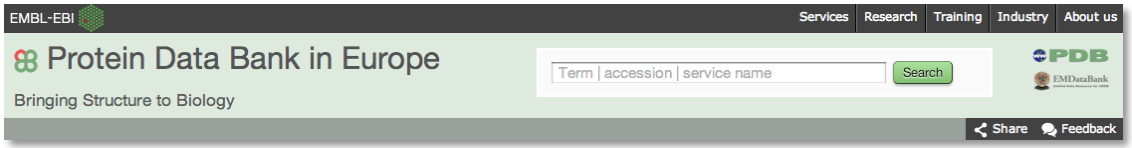


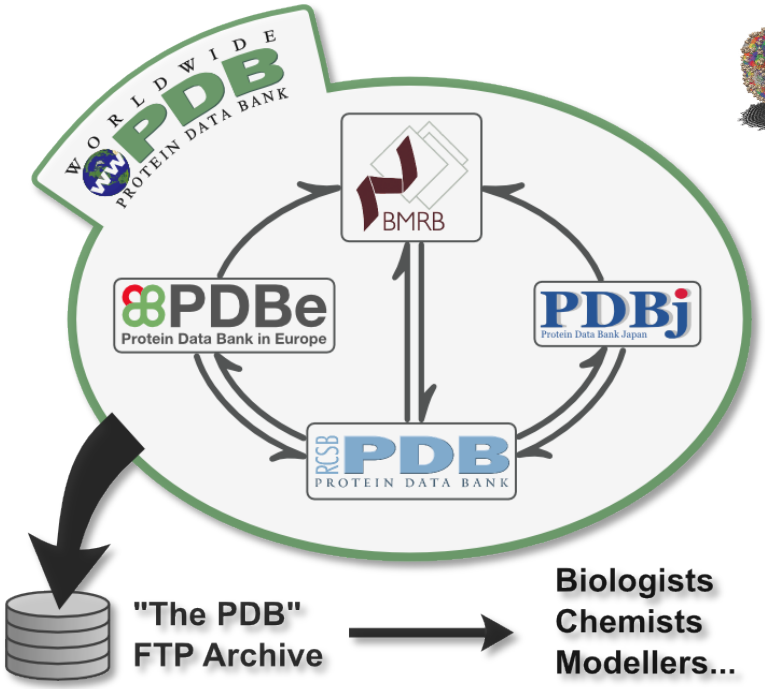


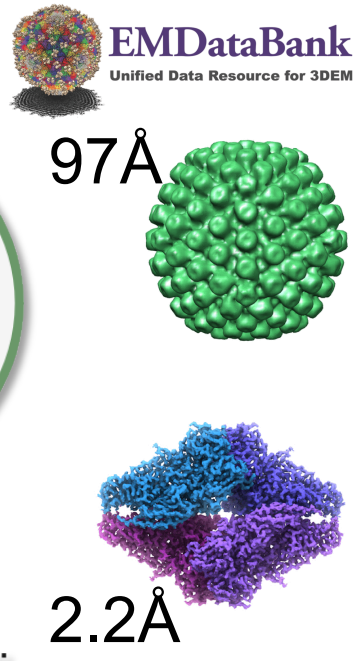




PDBe at a glance







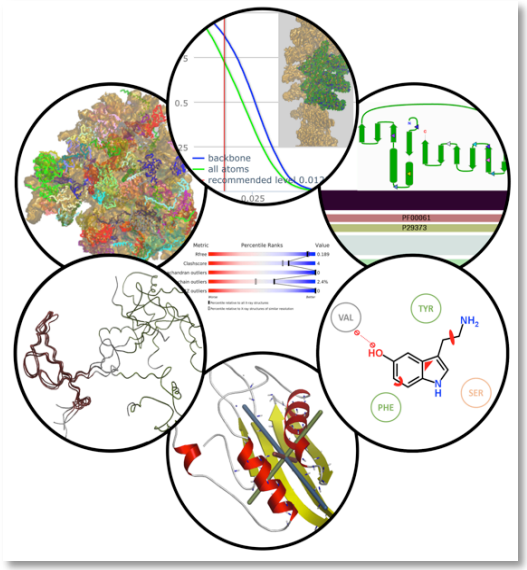
Major activities:

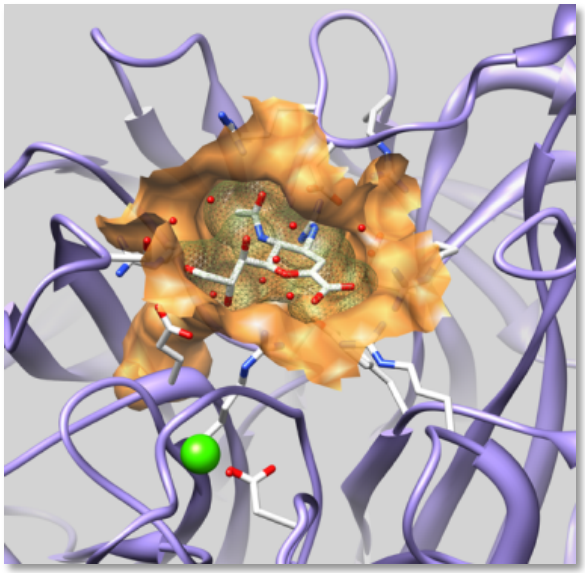
- Deposition and annotation site for structural data on biomacromolecules & complexes (X-ray, NMR, EM)
- Integrated resource to serve structural data and information

Mission: Bringing structure to biology

Bringing structure to biology

Molecular data → Information → Knowledge



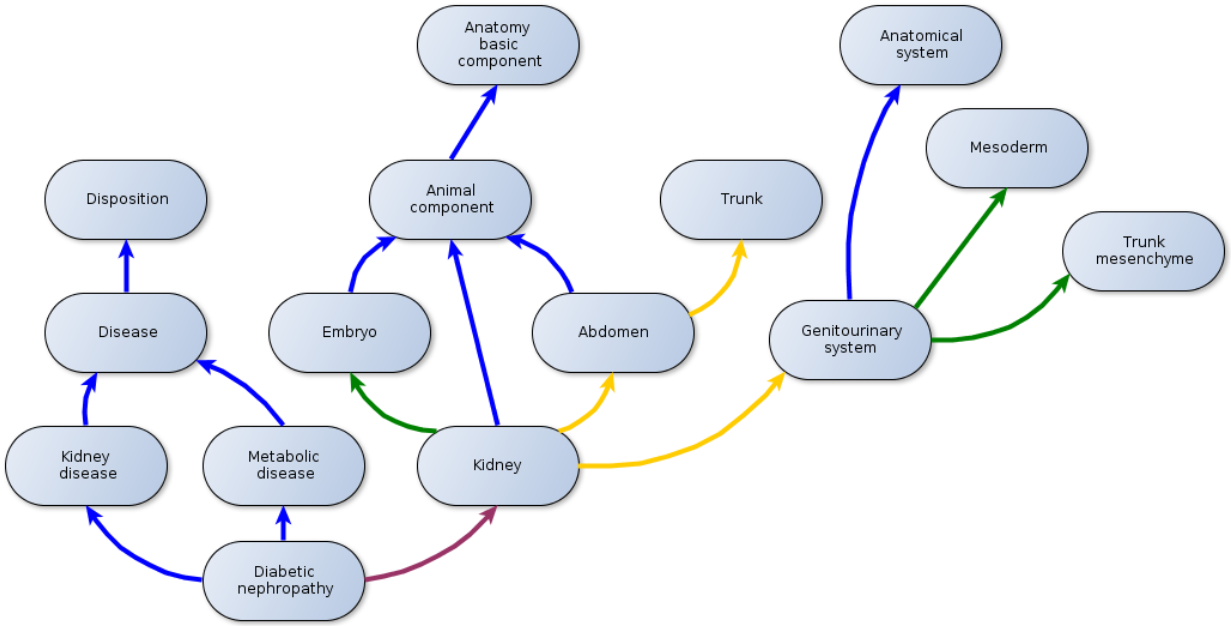


Mutation X disrupts enzyme function, which causes disease Y

"Coordinates by themselves just specify shape and are not necessarily of intrinsic biological value, unless they can be related to other information"

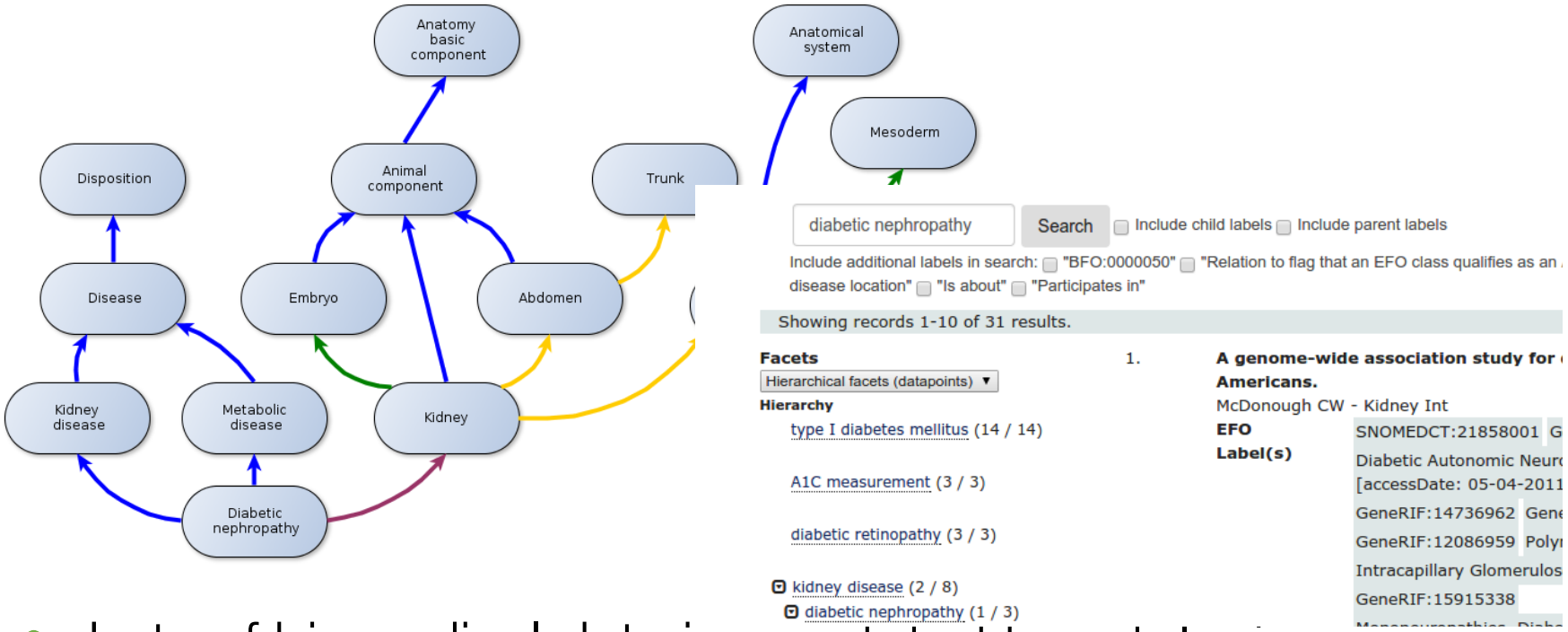
Integrative database analysis in structural genomics, Mark Gerstein, Nature Structural Biology 7, 960 (2000)

Case Study 1 – Faceting with Ontologies



- Lots of biomedical data is annotated to ontologies
- Ontologies can be used better to enrich search
  - Hierarchy, relations between concepts all possible
- Exploring dynamic, ontology enriched faceting in BioSolr

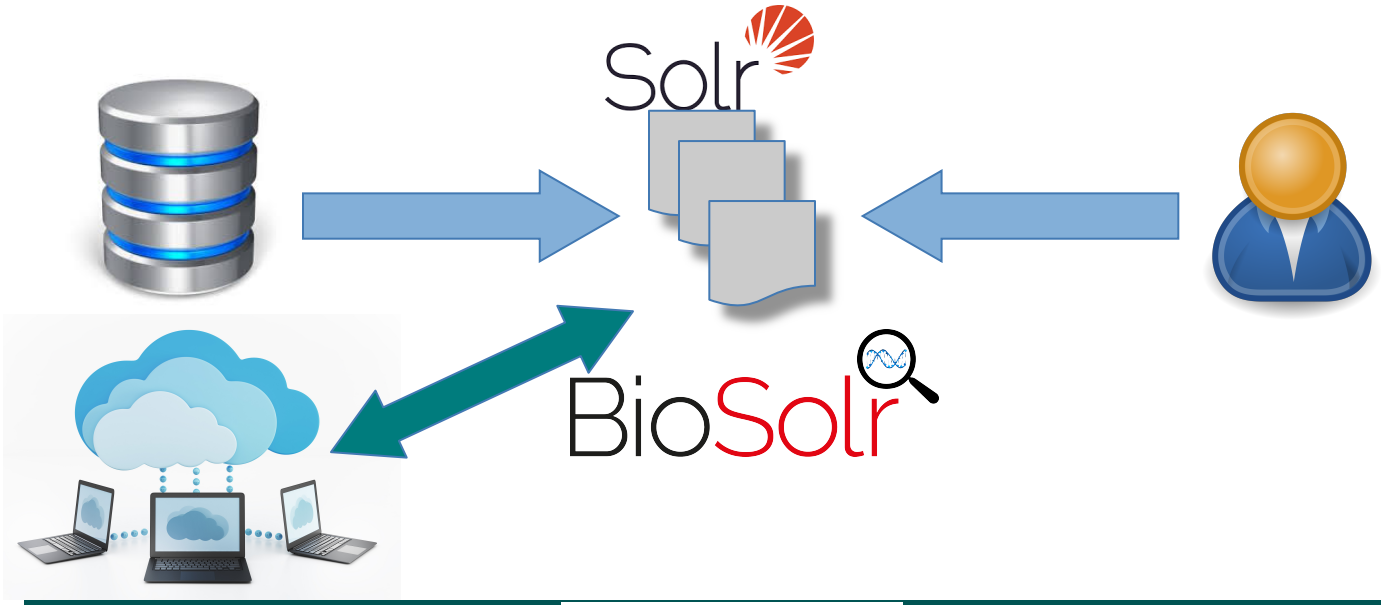
Case Study 1 – Faceting with Ontologies



- Lots of biomedical data is annotated to ontologies
- Ontologies can be used better to enrich search
  - Hierarchy, relations between concepts all possible
- Exploring dynamic, ontology enriched faceting in BioSolr

Case Study 2 – X-joins

- Mechanism for joining Solr indexes with external services
- Federate and embed external search results (e.g. BLAST, FASTA search) in returned documents
- Can The scores be integrated in Solr
  - boosts, filtering, nested queries and more



Other areas of interest

- Optimal indexing strategies
- Autocompletion
- Similarity searches
- Core joins
- Solving “megacores”
- Indexing hierarchical data (including ontologies)
- Faceting with ontologies
- Ngram highlighting
- Solr SPARQL integration
- Search federation and distribution
- Solr “best practices”
- Community building

Finding out more and getting involved...

https://github.com/flaxsearch/BioSolr

https://www.ebi.ac.uk/seqdb/confluence/display/BIOSOLR

www.flax.co.uk/blog

#biosolr, @PDBeurope, @flaxsearch

solr-users@ebi.ac.uk

Program 2nd Feb 2016

- Session 1: Setting the scene (10:00-12:00)
  - Recap of BioSolr project – Charlie Hull and Sameer Velankar
  - The State of Solr - Eric Pugh
  - BioSolr Developments - Tom Winch and Matt Pearce, Flax
- Lunch break (12:00-1:00)
- Session 2: Solr Applications (1:00-2:30)
  - New Developments in Search at NCBI: Querying Feature Annotations & High Availability Solr Stack in AWS - Peter Meric, NCBI
  - Seqr: searching protein sequences by similarity in Solr - Lewis Geer, NCBI
  - Using Solr sharding to provide federated search for plant phenotypes? - Dan Bolser, Ensembl Genomes, EBI
  - Ensembl data: how far will Elasticsearch stretch? - Dan Staines, Ensembl Genomes, EBI

Program 2nd Feb 2016

- Tea/coffee break
- **Session 3: Hands-on workshop** (3:00-4:30) – Training room
- Bar at Hinxton Hall (6:00)
- Dinner (6:30)

Program 3rd Feb 2016

- Session 4: Elastic search and other search technologies (10:00-12:00)
  - The State of Elastic search - Eric Pugh
  - Elastic search developments in BioSolr - Matt Pearce, Flax
  - Kibi, an Open Source Kibana fork for relational/graph exploration & analytics - Giovanni Tummarello, Siren Solutions
  - The twisted path towards search scalability in chemical biology big data while on a budget– Evan Bolton, PubChem, NCBI
  - Using Elasticsearch to store, integrate and mine diverse biological data. - Andrea Pierleoni, CTTV
- Lunch break (12:00-1:00)

- Session 5: Search Applications (1:00-2:30)
  - Literature Services: Migrating from Lucene to Solr Nikos Marinos, Literature services, EBI
  - A personal experience developing systems requiring search functionality - Rafael Jimenez, Elixir
  - Using django-haystack to provide SOLR-based search capabilities for the ChEMBL REST API - Michal Nowotka, ChEMBL, EBI
  - EBISearch, search your biologial data at EMBL-EBI - Nicola Buso, Web production, EBI
- Tea/coffee break (2:30-3:00)
- Hands-on Workshop (3:00-4:30) – Training room
- Discussion (4:30-5:00)

Who are Flax?

- Building open source search applications since 2001
- Independent, honest **advice** and analysis
- Expert **design** & **development**, Apache Solr committers
- UK Authorized Partner of
- Test-driven relevancy and performance tuning
- Custom **training** & **mentoring**

ENTERPRISE SEARCH Cambridge

Lucidworks

Lucene Solr London User Group

BBSRC 20 Years of Pioneering Great British Bioscience

flax The Open Source Search Specialists

EMBL-EBI

What have we achieved?

- Facet.contains (now in Solr 5.1 and above - SOLR-1387)
- Searching across multiple external datasets (Xjoin)
  - See SOLR-7341 (and please up-vote!)
- Researched searching across multiple Solr nodes across different locations.
- Indexing ontologies – both with and without document data.
- Enriching documents with ontology data.
  - and facet trees
- ...also various ad-hoc advice on best practices for indexing, scaling and Solr use generally!

And more achievements!

- User group on campus – 3 meetings, some with external speakers
- Talked about BioSolr at the Lucene/Solr London Usergroup
- Presented at Lucene Revolution 2015, Austin, Texas
- Poster session at BOSC 2015, Dublin, Ireland
- Tutorial at SWAT4LS 2015, Cambridge
- Enquiries and downloads from other institutions
- Many blog posts on BioSolr, see
- New PDBe Search uses BioSolr improvements

Get Involved!

- Check out the github page: https://github.com/flaxsearch/BioSolr
- Vote for Xjoin: https://issues.apache.org/jira/browse/SOLR-7341
- Suggest more use cases for what we've built!

Thank you for listening – any questions?

www.flax.co.uk/blog

+44 (0) 8700 118334

Twitter: @flaxsearch

The Open Source Search Specialists