How can text mining scale to meet diverse and precise curation needs?

Jo McEntyre
Literature Services
www.ebi.ac.uk
@jomcentyre
Curation makes data FAIR

- Analyse, curate and integrate
- Tools to assist reuse
- Share with other data providers
- Deposition
- Citation, reproducibility
- Classification
- Collaborative enterprise, community standards
Data resources at EMBL-EBI

- **Literature & ontologies**
  - Experimental Factor Ontology
  - Gene Ontology
  - BioStudies
  - Europe PMC

- **Genes, genomes & variation**
  - Ensembl
  - Ensembl Genomes
  - GWAS Catalog
  - Metagenomics portal

- **Gene, protein & metabolite expression**
  - Expression Atlas
  - Metabolights
  - PRIDE
  - RNA Central

- **Protein sequences, families & motifs**
  - InterPro
  - Pfam
  - UniProt

- **Molecular structures**
  - Protein Data Bank in Europe
  - Electron Microscopy Data Bank

- **Chemical biology**
  - ChEBI
  - ChEMBL
  - SureChEMBL

- **Systems**
  - BioModels
  - BioSamples
  - Enzyme Portal
  - IntAct
  - Reactome

- **Molecular Archives**
  - European Nucleotide Archive
  - European Variation Archive
  - European Genome-phenome Archive
  - ArrayExpress
Open Targets

A partnership to transform drug discovery through the systematic identification and prioritisation of targets
Target = Gene or Protein (integrated through EnsEMBL gene ID)

Disease or Phenotype

Ontology (SPOT Team)

http://targetvalidation.org/
The Imaging Revolution

<table>
<thead>
<tr>
<th>1890</th>
<th>1940</th>
<th>1990</th>
<th>2040</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biochemistry</td>
<td>Molecular Biology</td>
<td>„omics“</td>
<td>Imaging</td>
</tr>
</tbody>
</table>

Virginia.tech, Genelon, Nature, EMBL
Towards defining reference data

Euro-BioImaging — ELIXIR DATA BRIDGE

Structural Imaging
Cellular Phenotyping
Cellular Atlas
Model Organism Atlas
Human Atlas

All images from Euro-BioImaging

All biomolecular data within ELIXIR
Need to integrate text mining with other approaches: data, engineering

• Our collaborative community facilitates social, scientific and technical interactions.

• This image shows internal interactions between data resources, as determined by the exchange of data.

• The width of each internal arc is weighted according to the number of different data types exchanged.
Beyond the PDF:
Infrastructure
1. Access to content

- Open archives such as PubMed, PMC, Europe PMC
  - Open terms of use
  - CC0
  - CC-BY
  - CC-BY-SA
  - … … …
2. Sharing Annotations

Platform A  Curators  Platform B
Professional  Community Curation

Mapping  Triage  Feedback  Crosslinking

Annotations

Infrastructural components

Content  Engineering  Metadata
• Text and data mining community.
• Providing deep links to data
Impediments to a biomedical text mining platform

- Current work practices by all stakeholders
- Entrenched business models
- Size of the undertaking aka resources needed
- Trust
- Incentives to use the platform


Phil Bourne
Annotations on SciLite

Background

Amyotrophic lateral sclerosis (ALS) is characterized by a progressive degeneration of motor neurons in brain and the spinal cord, resulting in muscle weakness. Patients eventually become paralyzed and approximately 50% die within 3 years of onset of symptoms, usually as the result of respiratory failure [1]. Although the precise mechanisms of ALS remain unclear, approximately 28% of patients with ALS have dominant mutations in the Cu/Zn superoxide dismutase 1 (SOD1) gene [2]. Transgenic mice overexpressing the mutant human SOD1 gene (mSOD1 mice) develop progressive motor neuron degeneration that resembles ALS and therefore these mice serve as an animal model for the disease [3].

Although ALS is characterized by motor neuron degeneration, infiltration of T lymphocytes are significant pathological features in patients and mSOD1 mice, and a role for these cells in the pathogenesis of ALS has been suggested. Recent experiments in mSOD1 mice suggest that neurons do not become autonomous and depend on the active participation of infiltrating T cells [7-9].

Microglia, resident immune effector cells in the central nervous system (CNS), display functional plasticity during activation, which involves changes in cell number, morphology, surface receptors, and production of growth factors and cytokines [10]. T-cell-derived cytokines play critical roles in the control of the microglial phenotype. For example, classically activated microglia (M1 microglia) differentiate in response to granulocyte macrophage colony-stimulating factor (GM-CSF) and are primed by interferon gamma (IFN-γ), one of the most important cytokines produced by T helper 1 (Th1) cells, in the presence of lipopolysaccharide (LPS) [10,11]. M1 microglia secrete increased proinflammatory cytokines, superoxide radicals, nitric oxide (NO), and reduced neurotrophic factors, which promote neuronal death [12]. In contrast, representative T helper 2 (Th2) cytokines, such as interleukin 4 (IL-4) and interleukin 13 (IL-13), can convert microglia, primed by macrophage colony-stimulating factor (M-CSF), to an alternatively activated M2 phenotype [12]. M2 microglia are also characterized by increased expressions of arginase 1 (Arg1), resistin-like alpha (Retnla), and chitinase 3-like 3 (Ym1), which play important roles in tissue repair and remodeling [10]. However, the precise roles of crosstalk between T cells and microglia in the pathology of ALS remain unknown.
Linking annotations in SciLite

Sentence level Link-back

Link-out
Annotations API

- API users can retrieve by
  - Annotation type (gene, disease, organism)
  - PMCID/PMID
  - Providers
  - Specific entity e.g. “heart attack” or “human”

- CC-0, CC-BY, CC-BY-NC plus “OA subset”

- Will be used in F11 Hackathon next week!
Acknowledgements

SciLite team:

• Johanna McEntyre – Team leader
• Jee-Hyub Kim – Text Miner
• Francesco Talo – Web developer
• Senay Kafkas – Text Miner
• Shaojiang Cai – Web developer
• Anusha Iyer - UX Architect
• Michelle-Ide Smith – Product Manager

Collaborators:

• HES-SO / SIB, Geneva:
  • Patrick Ruch
  • Julien Gobeil

• IntAct:
  • Sandra Orchard
  • Pablo Porras

• DisGeNET:
  • Laura Furlong

Europe PMC
FAIR data, open data

Findable
Accessible
Interoperable
Reusable

“Generic” data citation helps
More discipline- and data-specific

Computation e.g. search algorithms