Advancing Discovery Science
Predictive, Evidential and Meta Analytical Methods

Michel Dumontier
Associate Professor of Medicine
Stanford Center for Biomedical Informatics Research
Stanford University

AAAI Fall Symposium
Accelerating Science: A Grand Challenge for AI
November 18, 2016
New discoveries are being made by “research parasites” using other people’s data

A common rejection module (CRM) for acute rejection across multiple organs identifies novel therapeutics for organ transplantation

Khatri et al. JEM. 210 (11): 2205
DOI: 10.1084/jem.20122709
towards automated knowledge discovery
Development of an intelligent system for scientific inquiry using the totality of web-accessible data and services.
Challenge

Efficient and uniform access to distributed, versioned and self-describing data and services for reproducible analyses
A fundamental inability to easily query and mine structured knowledge
Linked Open Data uses the web as a framework to share and link semantically annotated data.
Descriptions of the data (metadata) are often messy, incomplete or missing.
Vast numbers of schemas and terminologies make for a confusing set of choices.
Making it Easier, Possibly Even Pleasant, to Author Interoperable Experimental Metadata

metadatacenter.org

NIH COMMONS

ImmPort studies
HIPC requirements

IBS sharing
ISA studies

SDR data

Yale University
Yale formalized assays
smartAPI: semantic and self describing REST APIs

OpenLifeData API
The OpenLifeData API to get network data
Version 1.0.0
Contact information
OpenLifeData API team
openlifedata-developer@googlegroups.com
http://openlifedata.org
Terms of service
http://openlifedata.org/terms/
License
MIT

Paths
/get/describe/{id}

Description
Returns data about a single URI

Parameters

<table>
<thead>
<tr>
<th>Name</th>
<th>Located in</th>
<th>Description</th>
<th>Required</th>
<th>Schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>path</td>
<td>ID of resource to lookup</td>
<td>Yes</td>
<td>string</td>
</tr>
</tbody>
</table>

Responses

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th>Schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>describe response</td>
<td></td>
</tr>
</tbody>
</table>

Try this operation
Field auto-suggestion w/conformance

```
swagger: '2.0'

info:
  version: "2.0.0"
  title: MyGene.info API
  description: Documentation of the MyGene.info Gene Query web services.
  contact:
    responsibleOrganization: The Scripps Research Institute
    responsibleDeveloper: The Scripps Research Institute
  email: help@mygene.info
```
Value auto-suggestion calls the smartAPI field-specific suggestion service.
Unify API data with Linked Open Data

a) Simplified MyGene.info object with JSON-LD context

```json
{  
  "@type": "http://identifiers.org/ncbigene/",
  "@context": {  
    "_id": "@id",
    "name": "http://schema.org/name",
    "interpro": {  
      "@id": "http://identifiers.org/interpro/",
      "@type": "@id"
    },
    "description": "http://schema.org/description"
  },
  "_id": "1017",
  "symbol": "CDK2",
  "name": "cyclin-dependent kinase 2",
  "interpro": {  
    "_id": "IPR000719",
    "description": "Protein kinase-like domain"
  }
}
```

b) Transformed JSON-LD object with semantic URIs included

```json
{  
  "_id": "1017",
  "@type": "http://identifiers.org/ncbigene/",
  "http://schema.org/name": "cyclin-dependent kinase 2",
  "http://identifiers.org/interpro": {  
    "_id": "IPR000719",
    "http://schema.org/description": "Protein kinase-like domain"
  }
}
```
The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier […] Barend Mons

Affiliations | Contributions | Corresponding author

Scientific Data 3, Article number: 160018 (2016) | doi:10.1038/sdata.2016.18
Received 10 December 2015 | Accepted 12 February 2016 | Published online 15 March 2016

FAIR: Findable, Accessible, Interoperable, Re-usable

Applies to all digital resources and their metadata
Challenge

Data will always be described using different schemas and vocabularies. Does that *still* matter? Can we *automate* the integration of data?
Schema and vocabulary heterogeneity is a challenge for data retrieval and data mining.

Three ways to model the relationship between a protein and the volume it occupies.
New Methods for Data Integration

• Many elegant solutions for entity or concept mappings, but these only offer an incomplete solution when combined with schemas

• Need to learn robust transformation patterns
  – Subsumption, Similarity, Analogy, ML, Probability

• Evaluate these in the context of use cases
  – Query answering
  – Data mining
  – Prediction
Challenge

Can we *scale* the validation of interesting findings?
Most published research findings are false
- John Ioannidis, Stanford University

The Problem of Reproducibility in Scientific Research

• Non-reproducibility of rates of 65–89% in pharmacological studies and 64% in psychological studies.

• Problem of multiple testing in high-dimensional experiments. For gene expression analyses, 26 of 36 (72%) genomic associations initially reported as significant were found to be over-estimates of the true effect when tested in other datasets.

• Analytic focus has been on significance (P) values rather than effect size or independent verification.
Support and Gap Analysis using Open Data and Open Services

- HyQue is a platform for *knowledge discovery* that uses data retrieval coupled with contradiction-based automated reasoning to validate scientific *hypotheses*
- Leverages *semantic technologies* to provide access to linked data, ontologies, and semantic web services
- Uses *positive and negative findings*, captures *provenance*
- Weighs *evidence* according to context
- Used to find *aging genes* in worm, assess *cardiotoxicity* of tyrosine kinase inhibitors


Text co-mentions
Gene ontology annotations
Differential gene expression

Access to open data
Linked to ontologies
Represented with a universal language
Queried using a portable language
Results stored with their provenance

Table 3 8 C. elegans genes that received the highest HyQue evaluations for their role in aging, the PubMed identifiers of papers describing their roles in regulating longevity, and the data evaluation functions that contributed to their scores

<table>
<thead>
<tr>
<th>WormBase identifier</th>
<th>Symbol</th>
<th>Score</th>
<th>PMID</th>
<th>Satisfied data evaluation function</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBGene000002805</td>
<td>sams-1</td>
<td>0.89</td>
<td>16103914</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>WBGene00000371</td>
<td>cco-1</td>
<td>0.78</td>
<td>21215371</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>WBGene000009741</td>
<td>dnr-1</td>
<td>0.78</td>
<td>16103914</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>WBGene000002178</td>
<td>jnk-1</td>
<td>0.78</td>
<td>15767565</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>WBGene000004013</td>
<td>pha-4</td>
<td>0.78</td>
<td>19239417</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>WBGene000004789</td>
<td>sgk-1</td>
<td>0.78</td>
<td>15068796</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>WBGene000004800</td>
<td>sir-2.1</td>
<td>0.78</td>
<td>21938067</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>WBGene000006796</td>
<td>unc-62</td>
<td>0.78</td>
<td>17411345</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
</tbody>
</table>

Table 4 31 highest scoring C. elegans genes that received HyQue evaluation scores for their role in aging without existing aging-related annotations, and the data evaluation functions that contributed to their scores

<table>
<thead>
<tr>
<th>WormBase identifier</th>
<th>Symbol</th>
<th>Score</th>
<th>PMID</th>
<th>Satisfied data evaluation function</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBGene000000252</td>
<td>bli-2</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene000000255</td>
<td>bli-5</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene000000262</td>
<td>bra-1</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene00000479</td>
<td>cgh-1</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene00000915*</td>
<td>daf-21</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene00001165</td>
<td>efn-4</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene00001428*</td>
<td>fkb-3</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene00001543*</td>
<td>gcy-18</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene00001578</td>
<td>ges-1</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene00001746</td>
<td>gsk-3</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene00001824</td>
<td>hbl-1</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene00001974</td>
<td>hmg-4</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene00001979</td>
<td>hmp-2</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene00002005*</td>
<td>hsp-1</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene00002013*</td>
<td>hsp-12.6</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene00002069*</td>
<td>ikb-1</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene00002881</td>
<td>let-756</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene00003029</td>
<td>lin-44</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene00003058</td>
<td>lov-1</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene00003210</td>
<td>mel-28</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene00003473</td>
<td>mtl-1</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene00003497</td>
<td>mup-4</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene00003977*</td>
<td>pes-2.1</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene00004392</td>
<td>mrr-2</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene00004765</td>
<td>sel-8</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene00006789</td>
<td>unc-54</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
<tr>
<td>WBGene00007036</td>
<td>sod-5</td>
<td>✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔</td>
<td></td>
</tr>
</tbody>
</table>
Scaling Validation

• Automated experimentation (Adam & Eve)
• Crowdsourcing
  – As a simple task
  – As an open problem
• Automated discovery of viable methods
• Automated implementation of viable methods
Key Research Challenges

• Scalable, shared, fault-tolerant, and readily re-deployable frameworks for archiving and providing versioned and maximally FAIR biomedical (meta)data
• Scalable methods for the prospective and retrospective authoring, assessment, and repair of metadata.
• Scalable methods to learn equivalent representational patterns
• Scalable frameworks for open, transparent, reproducible and recurrent analysis and meta-analysis of FAIR research data.
• Methods to identify investigative biases and knowledge gaps
• Scalable and reliable methods for the prioritization scientific hypotheses using evidence gathered across scales and sources
• Scalable methods for validation of research findings.