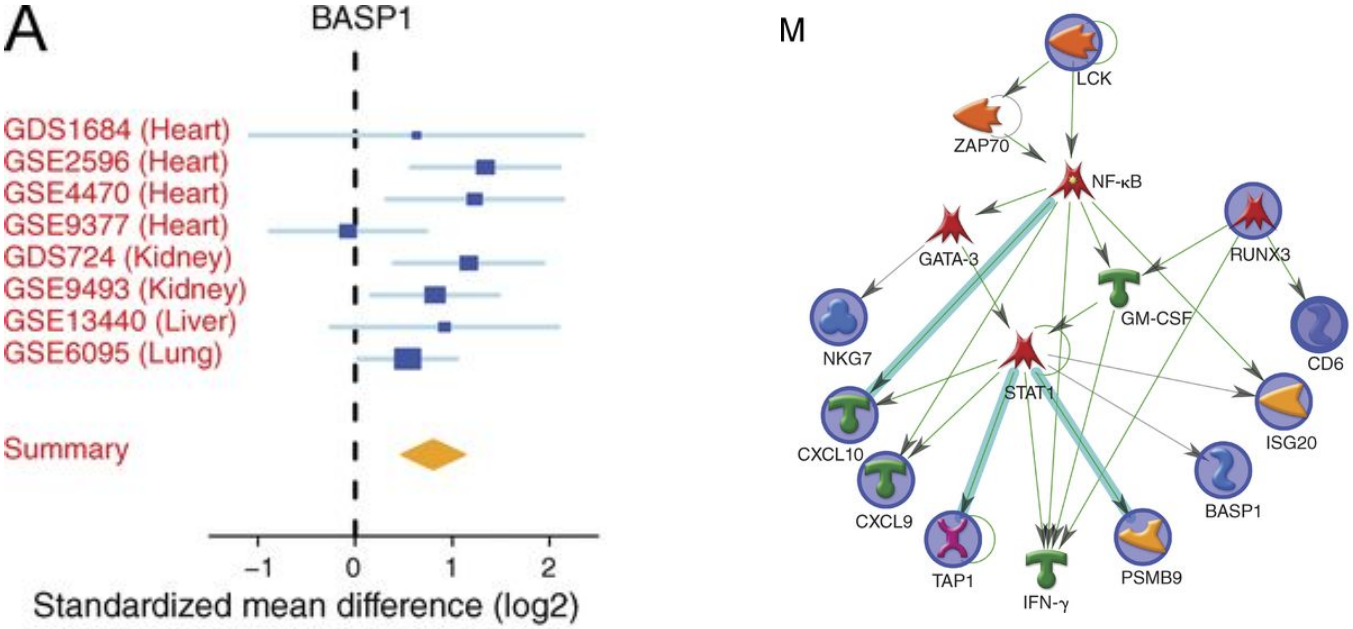
Advancing Discovery Science
Predictive, Evidential and Meta Analytical Methods

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

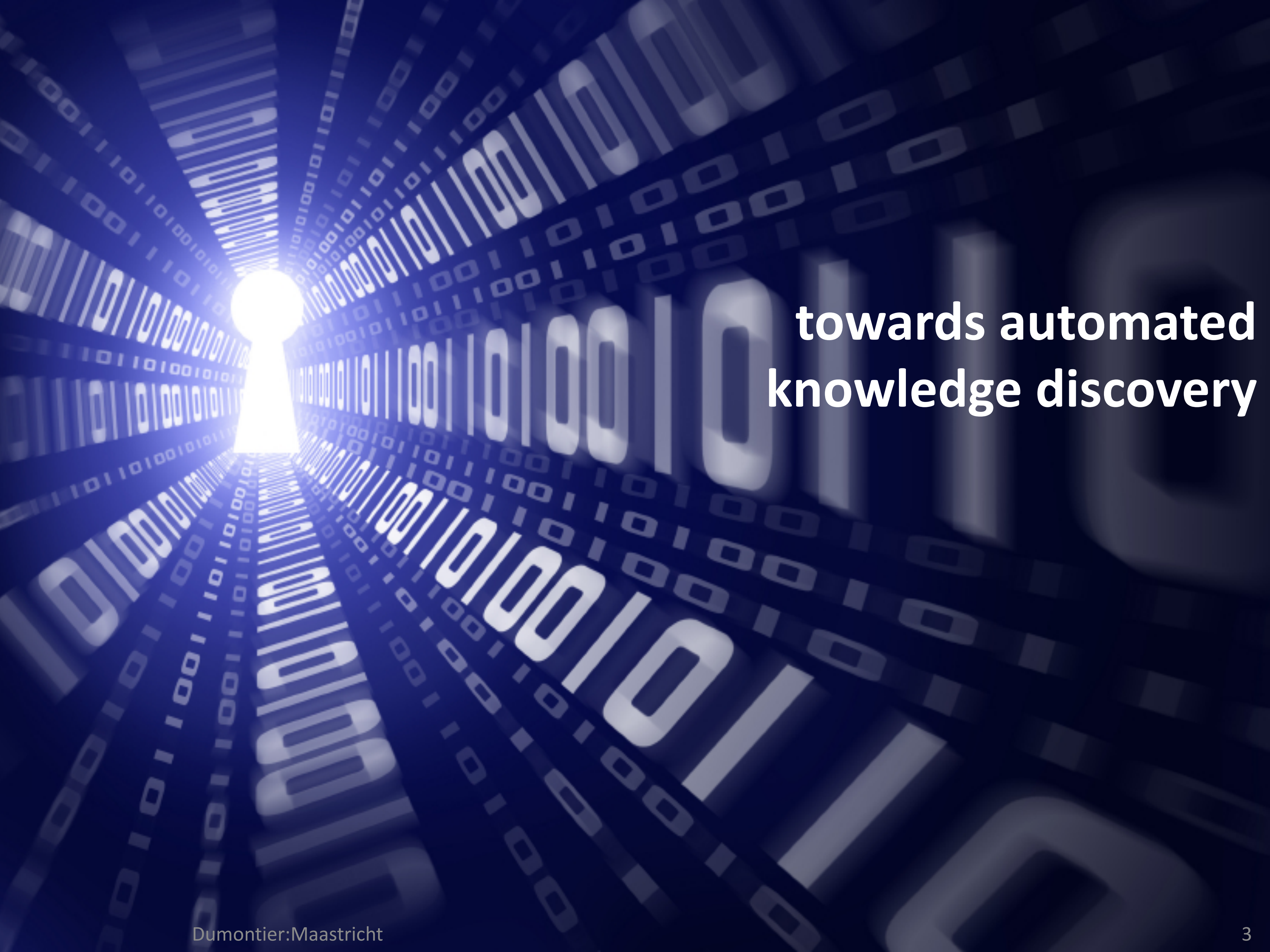
New discoveries are being made by “research parasites” using other people’s data



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 <u>totality of web-accessible data and services</u>.

Challenge

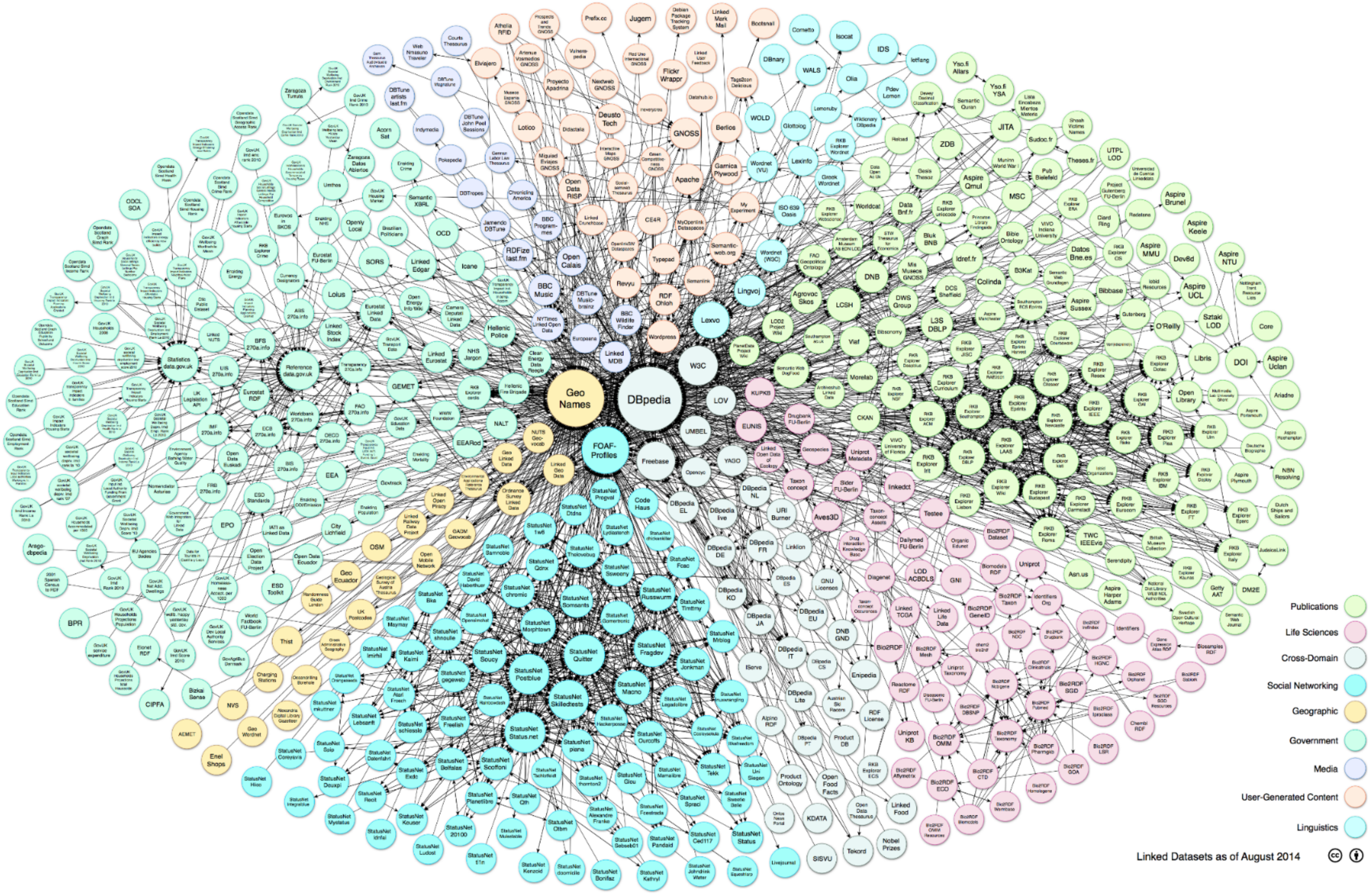
Efficient and uniform **access** to *distributed, versioned* and *self describing* ***data*** and ***services***

for *reproducible* **analyses**

A fundamental inability to easily query and mine structured knowledge

prosite
SIB
GO
the Gene Ontology
ChEBI
UniProt
the universal protein resource
UniSTS
Integrating Markers and Maps
PubMed
www.p
Entrez Gene
RCSB
PDB
PROTEIN DATA BANK
BIOCARTA
PubChem
KEGG
Kyoto Encyclopedia of Genes and Genomes
OBO
open biomedical ontologies
Pfam
UniParc
The UniProt Archive
MGI
MeSH
OMIM
HG
NC
KEGG COMPOUND
Taxonomy Browser
UniRef
Non-redundant Reference
100%>90%>50%
Entrez
Nucleotide
ProDom
InterPro
Protein Archive
HomoloGene
Discover Homologs

Linked Open Data uses the web as a framework to share and link semantically annotated data



BIO2RDF

BioModels

BioSamples

ChEMBL

Ensembl

Expression Atlas

Reactome

UniProt

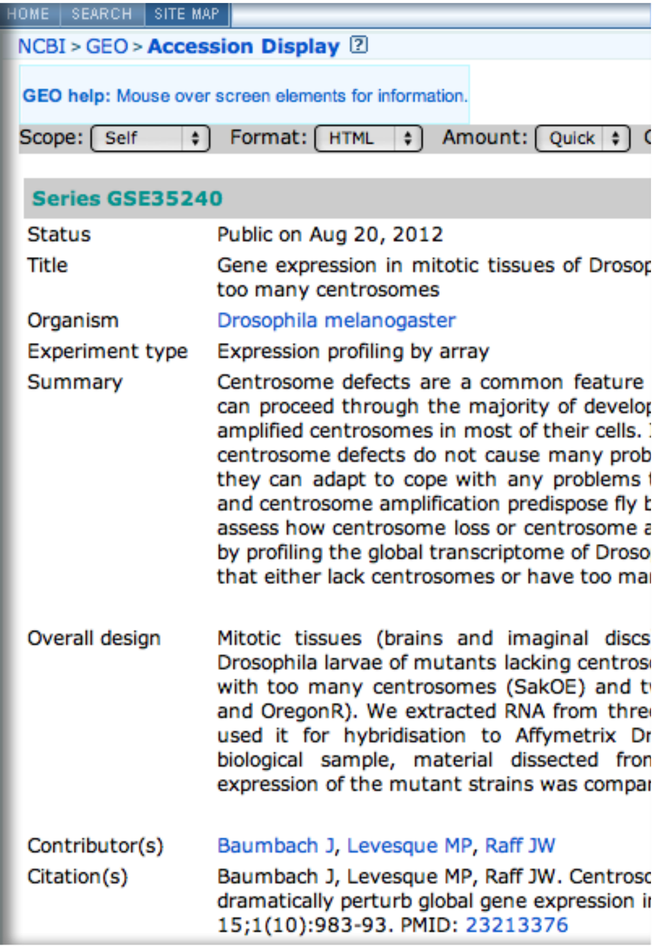
KEGG Kyoto Encyclopedia of Genes and Genomes

PubChem

MeSH

W3C Semantic Web

Descriptions of the data (metadata) are often messy, incomplete or missing



NCBI > GEO > **Accession Display**

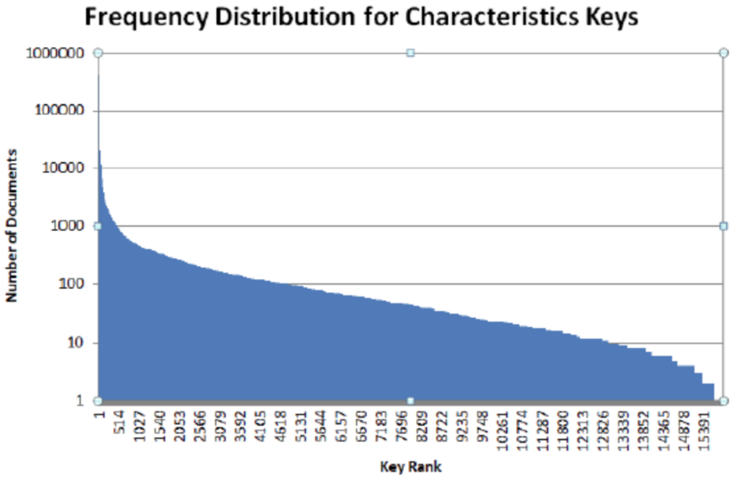
GEO help: Mouse over screen elements for information.

Scope: Self Format: HTML Amount: Quick

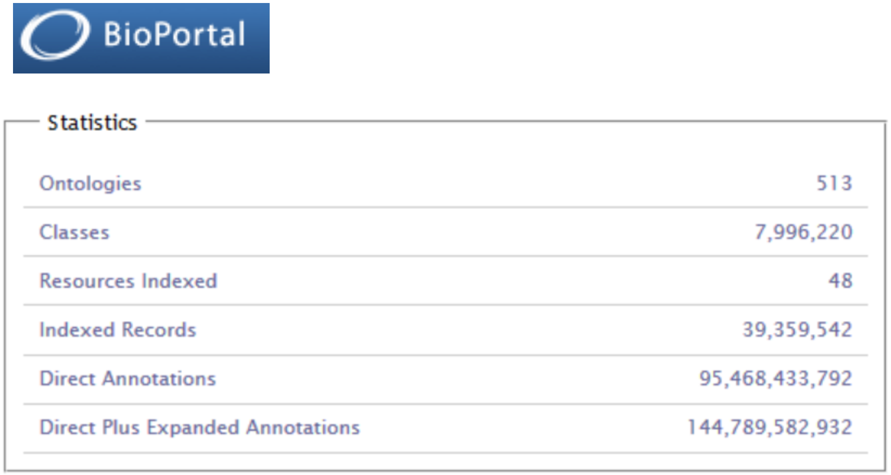
Series GSE35240

Status	Public on Aug 20, 2012
Title	Gene expression in mitotic tissues of Drosop too many centrosomes
Organism	Drosophila melanogaster
Experiment type	Expression profiling by array
Summary	Centrosome defects are a common feature can proceed through the majority of develop amplified centrosomes in most of their cells. centrosome defects do not cause many prob they can adapt to cope with any problems and centrosome amplification predispose fly b assess how centrosome loss or centrosome a by profiling the global transcriptome of Droso that either lack centrosomes or have too ma
Overall design	Mitotic tissues (brains and imaginal discs Drosophila larvae of mutants lacking centros with too many centrosomes (SakOE) and t and OregonR). We extracted RNA from thre used it for hybridisation to Affymetrix Dr biological sample, material dissected fron expression of the mutant strains was compar
Contributor(s)	Baumbach J, Levesque MP, Raff JW
Citation(s)	Baumbach J, Levesque MP, Raff JW. Centros dramatically perturb global gene expression i 15;1(10):983-93. PMID: 23213376

age	207147
Age	18089
age (yrs)	9891
age (years)	9272
age (y)	6226
age in years	1387
age_years	607
AGE	588
age(years)	558
age (year)	433
age (yr)	373
Age (years)	318
age (in years)	310
Age(years)	267
age [year]	97
age [y]	84
age [years]	83
Age(yrs.)	81
age.year	70
age (yr-old)	64
age(yrs)	59
age of patient	40
Age, year	39
Age (yrs)	36
Age of patient	33
age, years	24
'Age	21
Age (Years)	20
age (after birth)	18
age, yrs	12
age of subjects	4



Vast numbers of schemas and terminologies make for a *confusing* set of choices

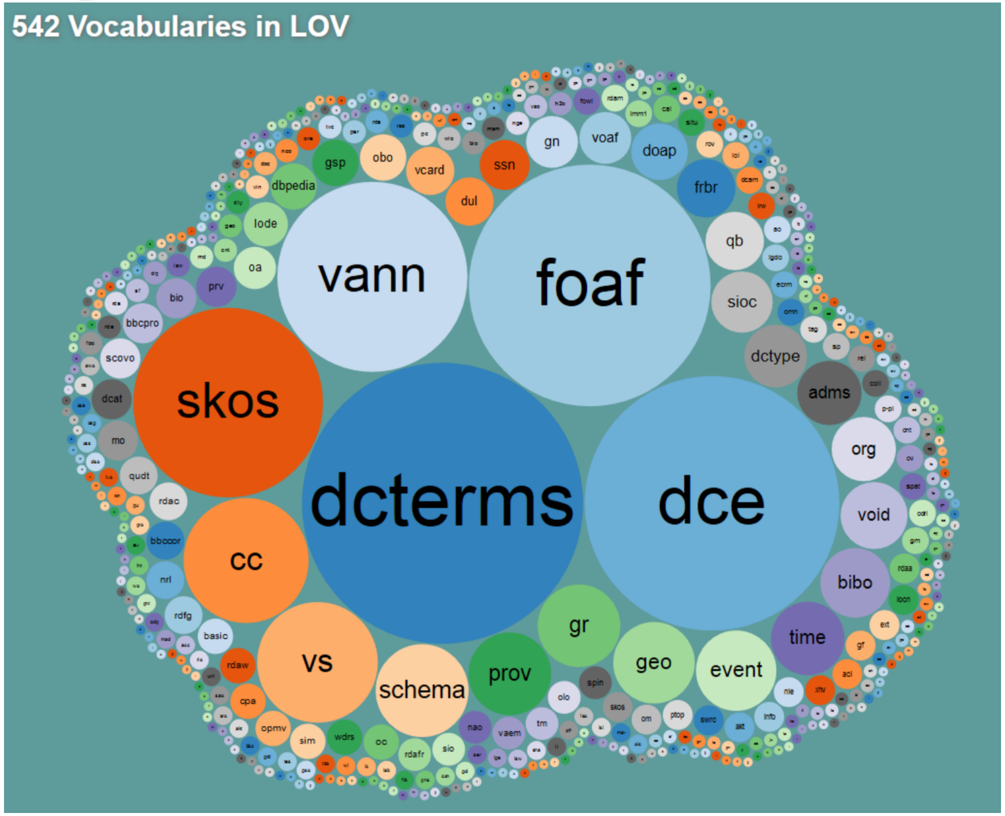


BioPortal

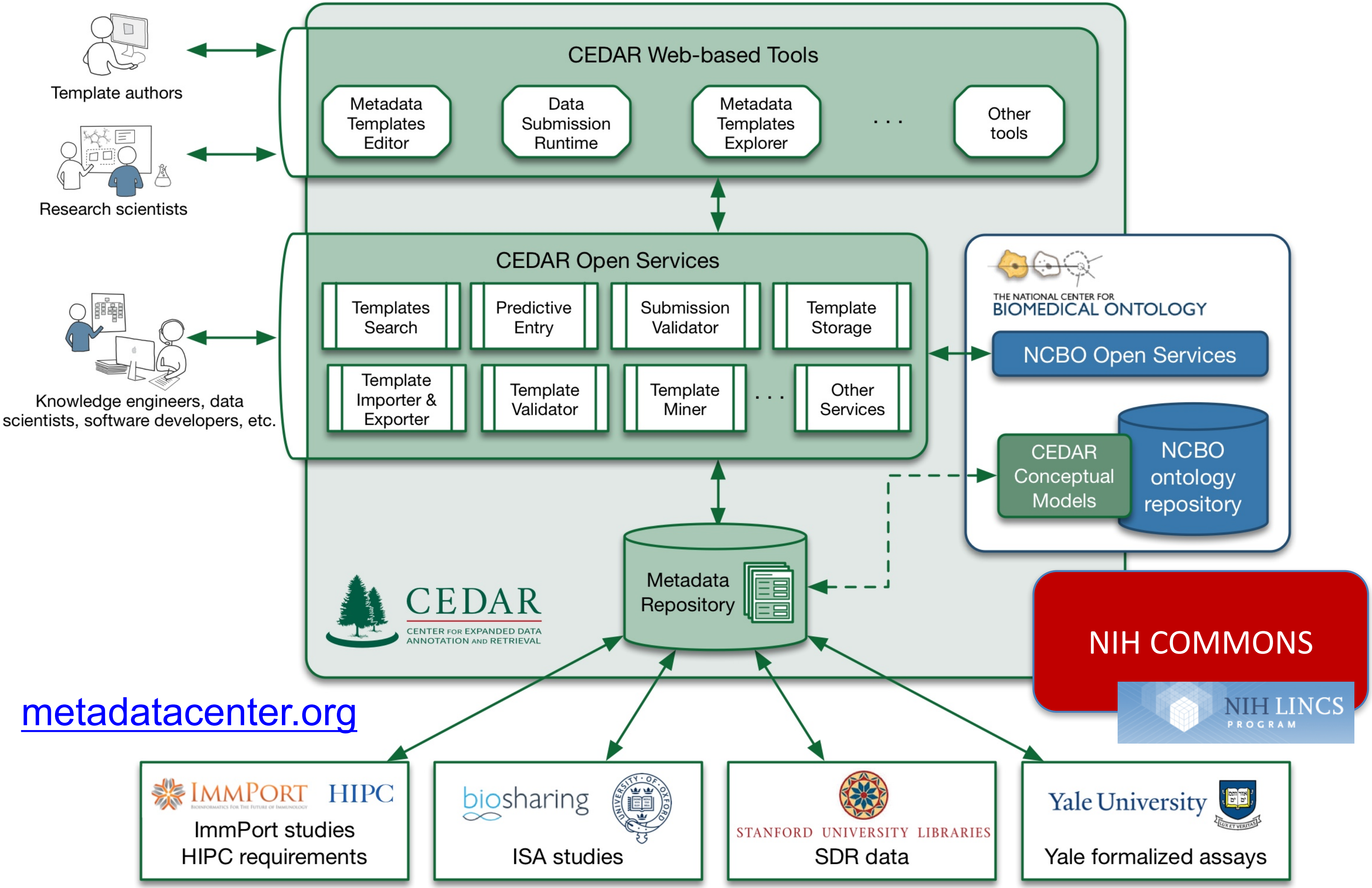
Statistics

Ontologies	513
Classes	7,996,220
Resources Indexed	48
Indexed Records	39,359,542
Direct Annotations	95,468,433,792
Direct Plus Expanded Annotations	144,789,582,932

Linked Open Vocabularies (LOV)

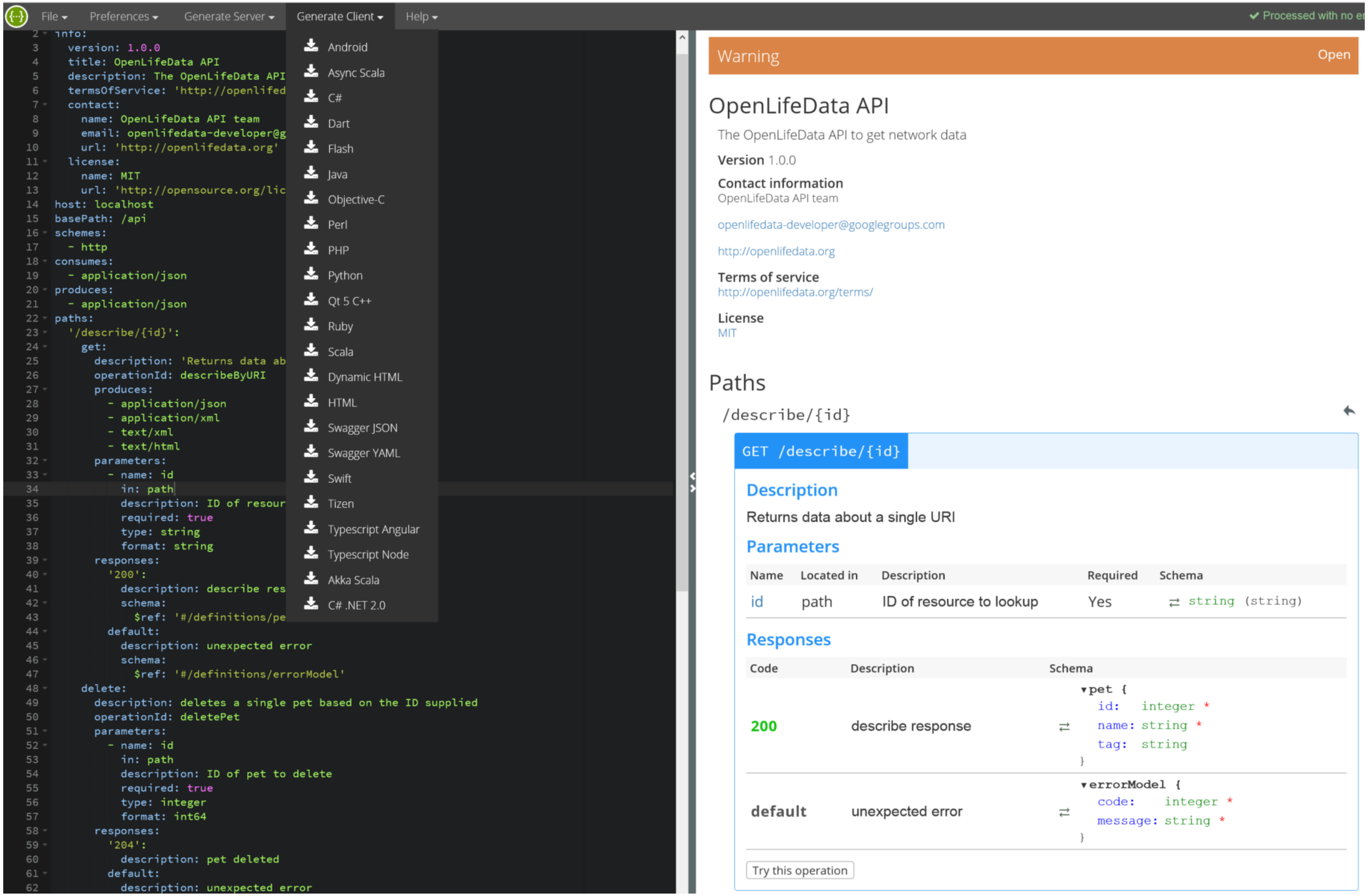


Making it Easier, Possibly Even Pleasant, to Author Interoperable Experimental Metadata



metadatacenter.org

smartAPI: semantic and self describing REST APIs



```
info:
  version: 1.0.0
  title: OpenLifeData API
  description: The OpenLifeData API
  termsOfService: 'http://openlifed
  contact:
    name: OpenLifeData API team
    email: openlifedata-developer@g
    url: 'http://openlifedata.org'
  license:
    name: MIT
    url: 'http://opensource.org/lic
host: localhost
basePath: /api
schemes:
  - http
consumes:
  - application/json
produces:
  - application/json
paths:
  '/describe/{id}':
    get:
      description: 'Returns data ab
      operationId: describeByURI
      produces:
        - application/json
        - application/xml
        - text/xml
        - text/html
      parameters:
        - name: id
          in: path
          description: ID of resour
          required: true
          type: string
          format: string
      responses:
        '200':
          description: describe res
          schema:
            $ref: '#/definitions/pe
        default:
          description: unexpected error
          schema:
            $ref: '#/definitions/errorModel'
    delete:
      description: deletes a single pet based on the ID supplied
      operationId: deletePet
      parameters:
        - name: id
          in: path
          description: ID of pet to delete
          required: true
          type: integer
          format: int64
      responses:
        '204':
          description: pet deleted
        default:
          description: unexpected error
```

Warning Open

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

/describe/{id}

GET /describe/{id}

Description

Returns data about a single URI

Parameters

Name	Located in	Description	Required	Schema
id	path	ID of resource to lookup	Yes	⇄ string (string)

Responses

Code	Description	Schema

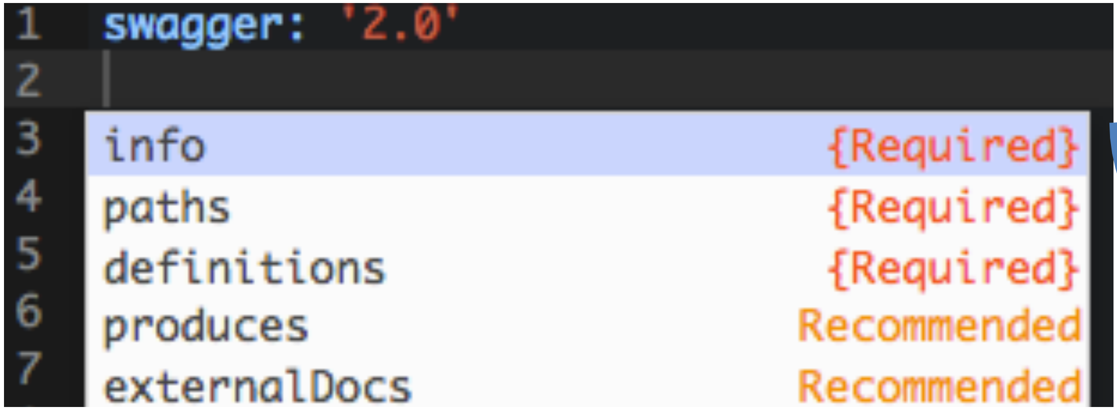
200	describe response	⇄ ▾pet {
		id: integer *
		name: string *
		tag: string
		}

default	unexpected error	⇄ ▾errorModel {
		code: integer *
		message: string *
		}

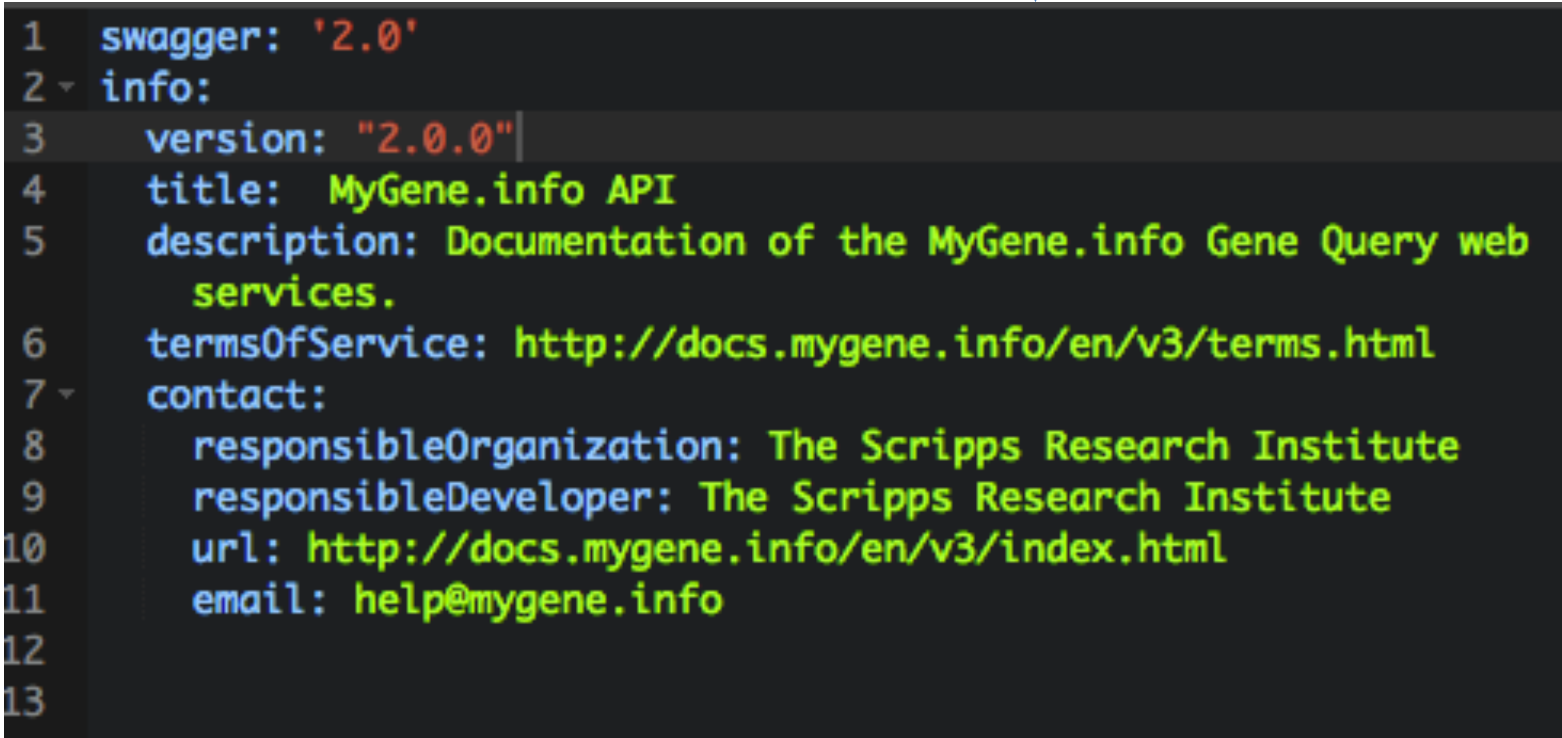
Try this operation

SWAGGER

Field auto-suggestion w/conformance



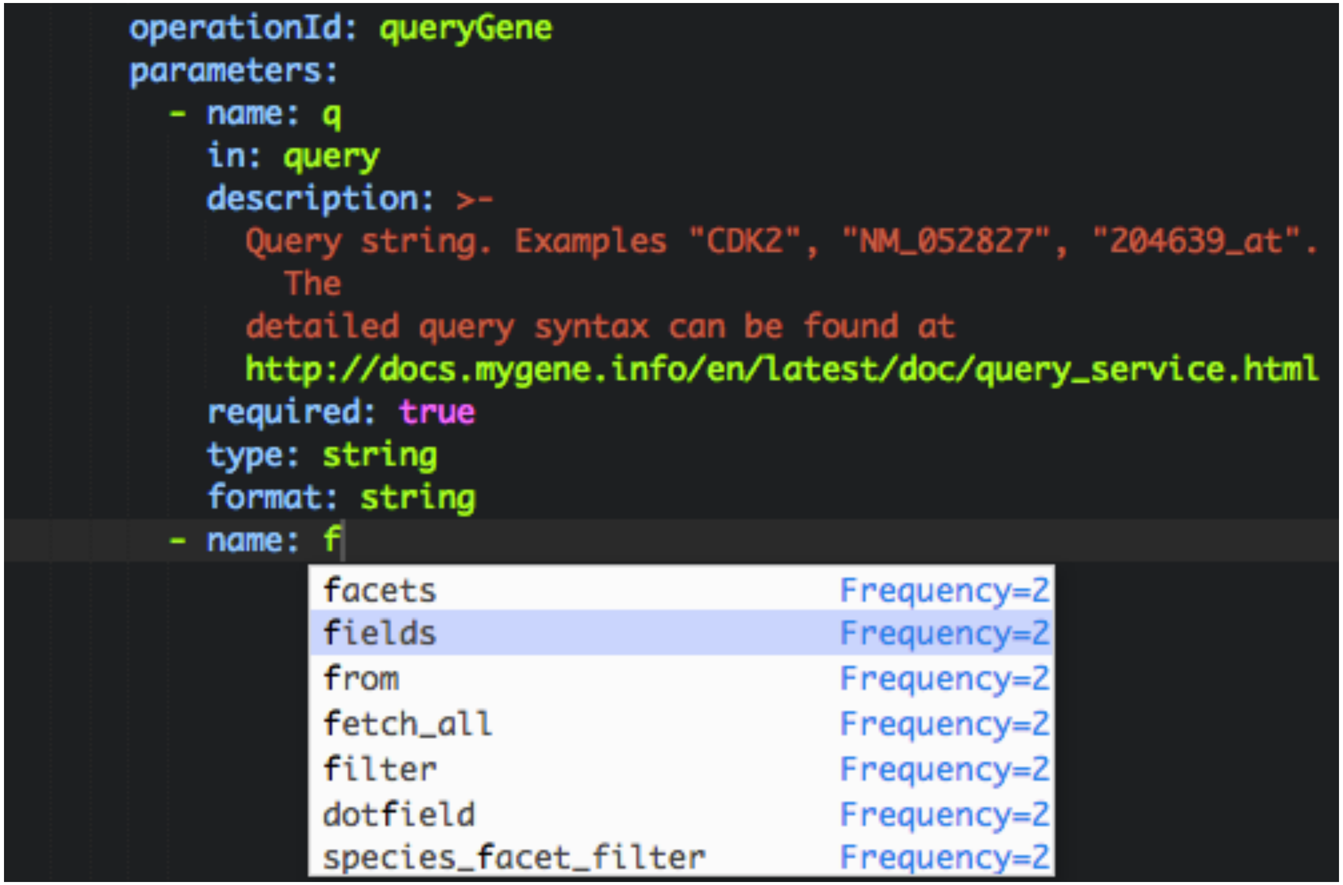
```
swagger: '2.0'

info                      {Required}
paths                     {Required}
definitions               {Required}
produces                  Recommended
externalDocs              Recommended
```



```
swagger: '2.0'
info:
  version: "2.0.0"
  title:  MyGene.info API
  description: Documentation of the MyGene.info Gene Query web services.
  termsOfService: http://docs.mygene.info/en/v3/terms.html
  contact:
    responsibleOrganization: The Scripps Research Institute
    responsibleDeveloper: The Scripps Research Institute
    url: http://docs.mygene.info/en/v3/index.html
    email: help@mygene.info
```

Value auto-suggestion

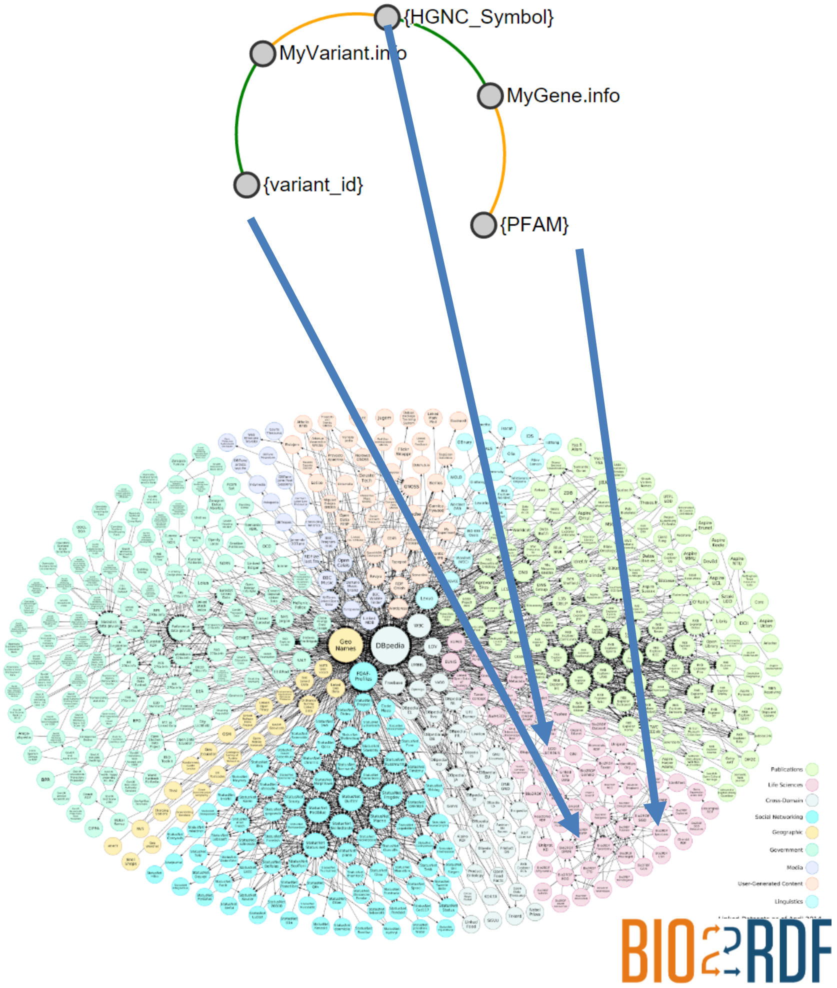
calls the smartAPI field-specific suggestion service

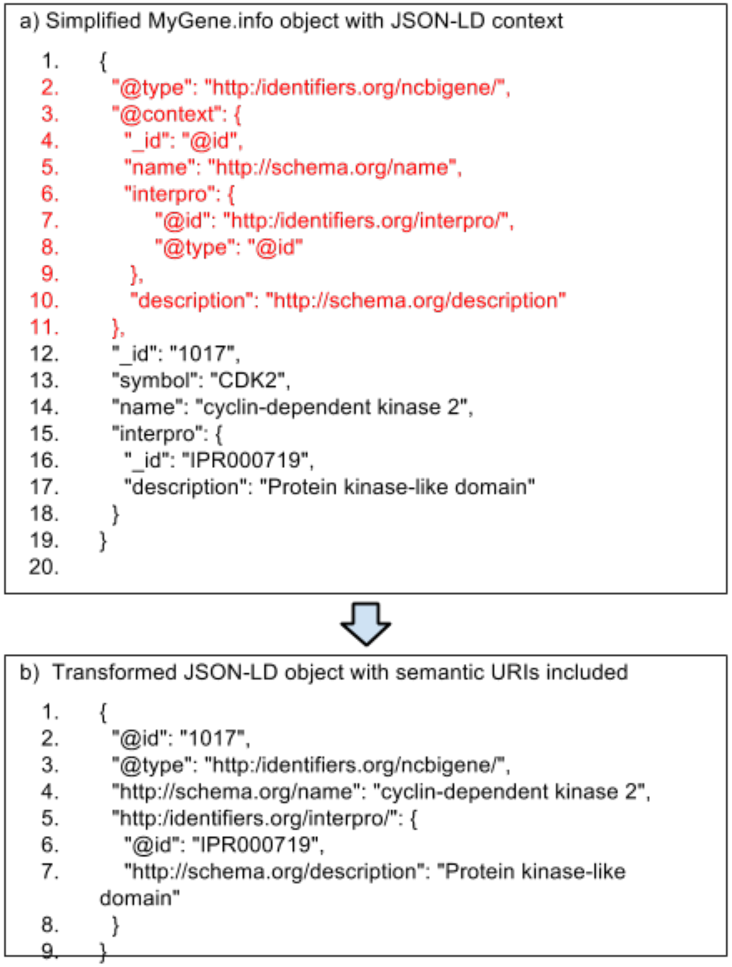


```
operationId: queryGene
parameters:
  - name: q
    in: query
    description: >-
      Query string. Examples "CDK2", "NM_052827", "204639_at".
        The
      detailed query syntax can be found at
      http://docs.mygene.info/en/latest/doc/query_service.html
    required: true
    type: string
    format: string
  - name: f
```

```
facets                          Frequency=2
fields                          Frequency=2
from                            Frequency=2
fetch_all                       Frequency=2
filter                          Frequency=2
dotfield                        Frequency=2
species_facet_filter            Frequency=2
```

Unify API data with Linked Open Data



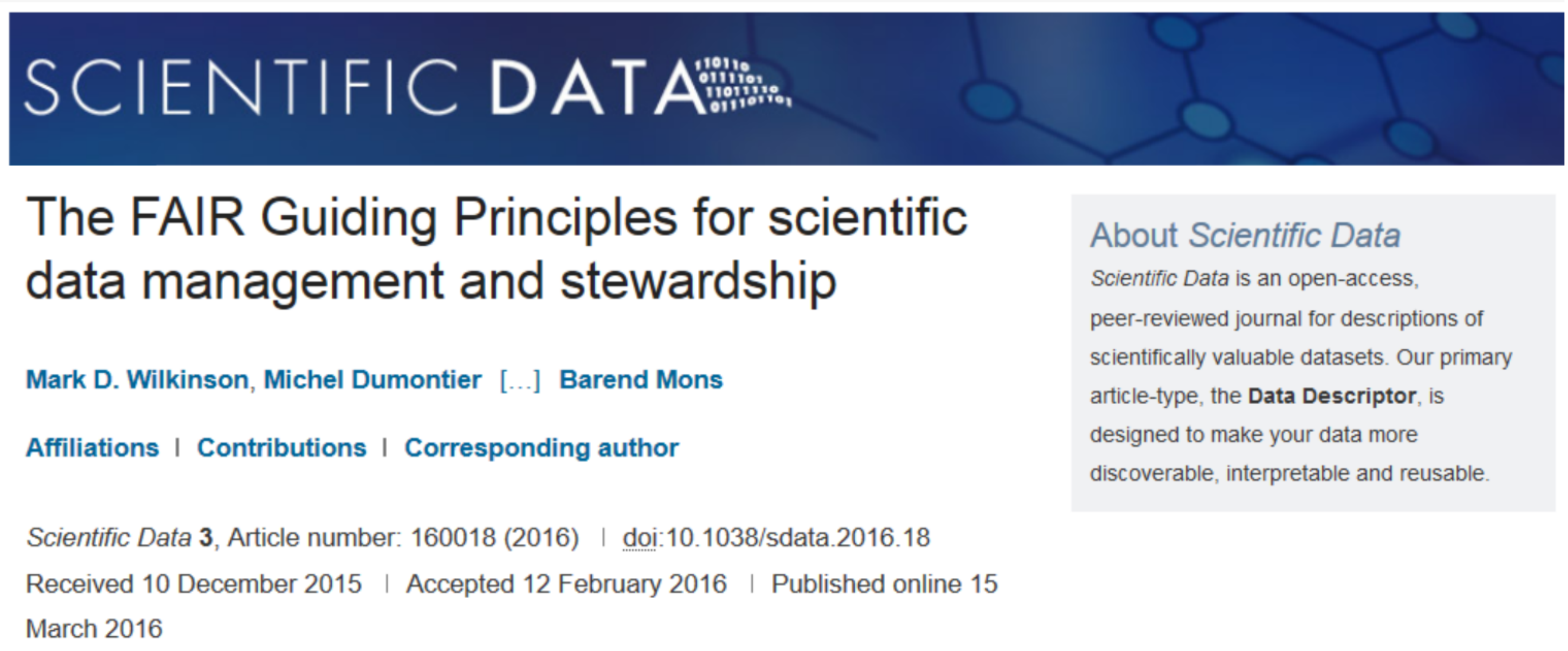


a) Simplified MyGene.info object with JSON-LD context

```
{
  "@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

```
{
  "@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"
  }
}
```



SCIENTIFIC DATA

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

About *Scientific Data*

Scientific Data is an open-access, peer-reviewed journal for descriptions of scientifically valuable datasets. Our primary article-type, the **Data Descriptor**, is designed to make your data more discoverable, interpretable and reusable.

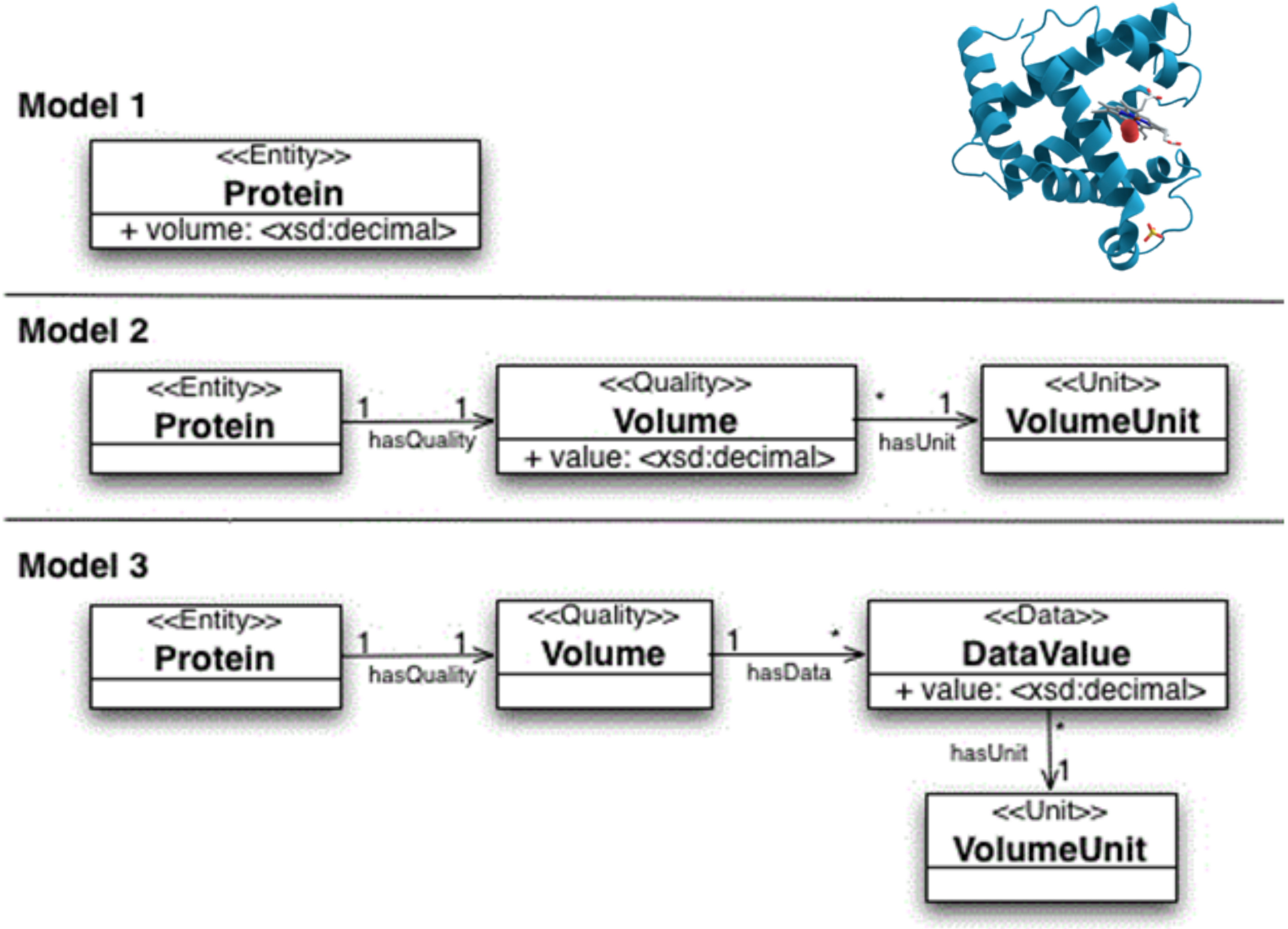
FAIR: Findable, Accessible, Interoperable, Re-usable

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 Mehtods 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

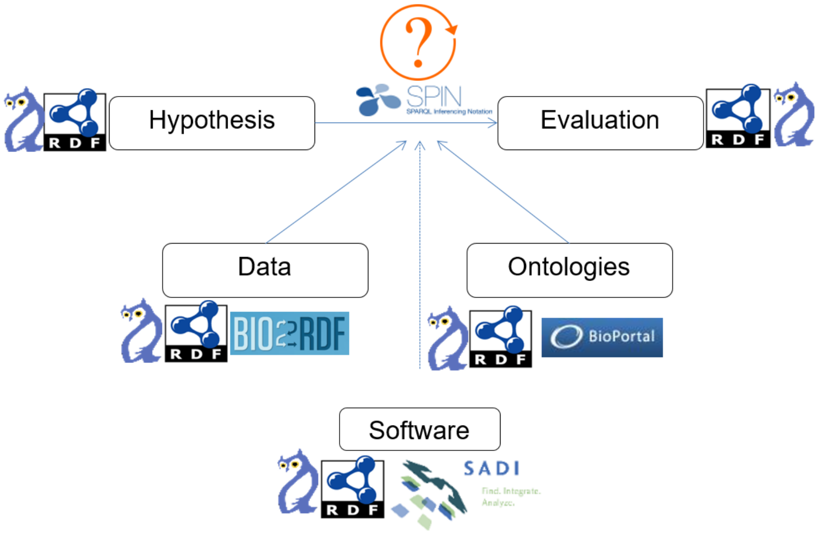
Ioannidis JPA (2005) Why Most Published Research Findings Are False. PLoS Med 2(8): e124.

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



HyQue: evaluating hypotheses using Semantic Web technologies. J Biomed Semantics. 2011 May 17;2 Suppl 2:S3.

Evaluating scientific hypotheses using the SPARQL Inferencing Notation. Extended Semantic Web Conference (ESWC 2012). Heraklion, Crete. May 27-31, 2012.

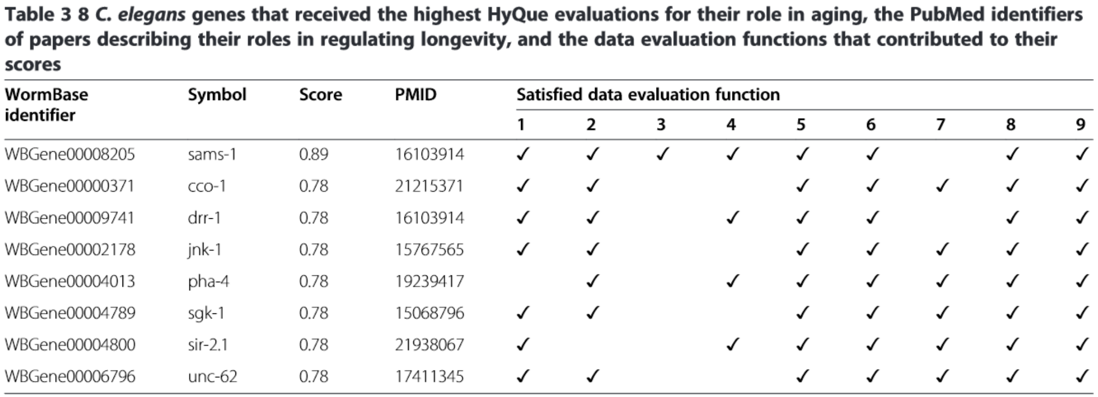


Table 3 8 *C. elegans* genes that received the highest HyQue evaluations for their role in aging, the PubMed identifiers of papers describing their roles in regulating longevity, and the data evaluation functions that contributed to their scores

WormBase identifier	Symbol	Score	PMID	Satisfied data evaluation function								
				1	2	3	4	5	6	7	8	9
WBGene00008205	sams-1	0.89	16103914	✓	✓	✓	✓	✓	✓		✓	✓
WBGene00000371	cco-1	0.78	21215371	✓	✓			✓	✓	✓	✓	✓
WBGene00009741	drr-1	0.78	16103914	✓	✓		✓	✓	✓		✓	✓
WBGene00002178	jnk-1	0.78	15767565	✓	✓			✓	✓	✓	✓	✓
WBGene00004013	pha-4	0.78	19239417		✓		✓	✓	✓	✓	✓	✓
WBGene00004789	sgk-1	0.78	15068796	✓	✓			✓	✓	✓	✓	✓
WBGene00004800	sir-2.1	0.78	21938067	✓			✓	✓	✓	✓	✓	✓
WBGene00006796	unc-62	0.78	17411345	✓	✓			✓	✓	✓	✓	✓

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

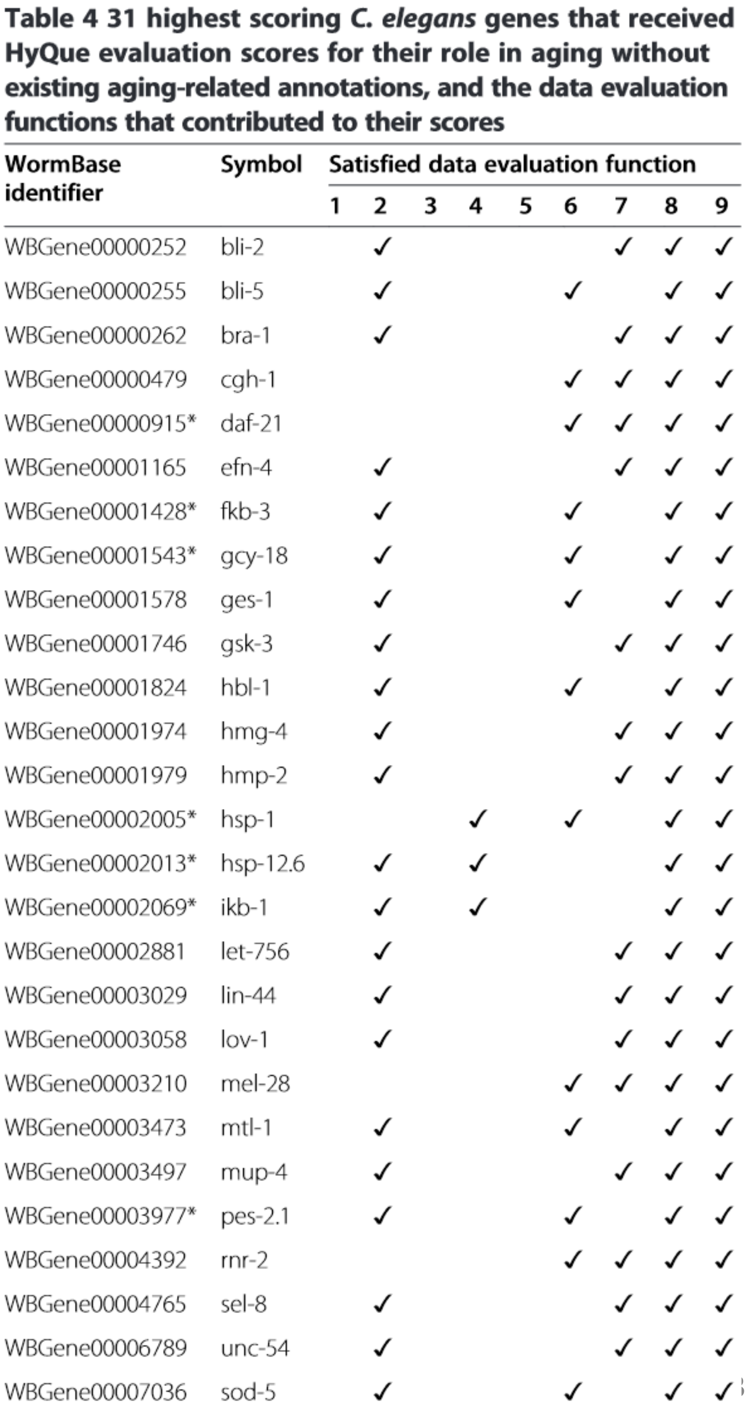


Table 4 31 highest scoring *C. elegans* genes that received HyQue evaluation scores for their role in aging without existing aging-related annotations, and the data evaluation functions that contributed to their scores

WormBase identifier	Symbol	Satisfied data evaluation function								
		1	2	3	4	5	6	7	8	9
WBGene00000252	bli-2		✓					✓	✓	✓
WBGene00000255	bli-5		✓				✓		✓	✓
WBGene00000262	bra-1		✓					✓	✓	✓
WBGene00000479	cgh-1						✓	✓	✓	✓
WBGene00000915*	daf-21						✓	✓	✓	✓
WBGene00001165	efn-4		✓					✓	✓	✓
WBGene00001428*	fkb-3		✓				✓		✓	✓
WBGene00001543*	gcy-18		✓				✓		✓	✓
WBGene00001578	ges-1		✓				✓		✓	✓
WBGene00001746	gsk-3		✓					✓	✓	✓
WBGene00001824	hbl-1		✓				✓		✓	✓
WBGene00001974	hmg-4		✓					✓	✓	✓
WBGene00001979	hmp-2		✓					✓	✓	✓
WBGene00002005*	hsp-1				✓		✓		✓	✓
WBGene00002013*	hsp-12.6		✓		✓				✓	✓
WBGene00002069*	ikb-1		✓		✓				✓	✓
WBGene00002881	let-756		✓					✓	✓	✓
WBGene00003029	lin-44		✓					✓	✓	✓
WBGene00003058	lov-1		✓					✓	✓	✓
WBGene00003210	mel-28						✓	✓	✓	✓
WBGene00003473	mtl-1		✓				✓		✓	✓
WBGene00003497	mup-4		✓					✓	✓	✓
WBGene00003977*	pes-2.1		✓				✓		✓	✓
WBGene00004392	rnr-2						✓	✓	✓	✓
WBGene00004765	sel-8		✓					✓	✓	✓
WBGene00006789	unc-54		✓					✓	✓	✓
WBGene00007036	sod-5		✓				✓		✓	✓

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.