identifiers.org

(Update and) Introduction

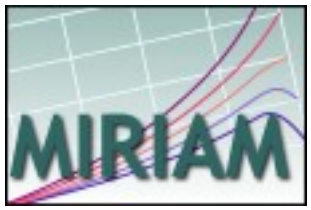
Nick Juty

COMBINE 2015
Salt Lake City
October 12th–16th

EMBL-EBI

MIRIAM guidelines (in a nutshell)

- Attribution information
  - Model name
  - Author information
  - terms of distribution
- Reference correspondence
  - publication reference
  - encoded model structure must reflect biological processes
  - reproduction of published results when instantiated in a simulation
- External data resources
  - annotation of model components using external resources (with URIs and qualifiers)



http://co.mbine.org/standards/miriam

Concept of annotation

model element + qualifier + (cross-)reference(s)

- **Cross-reference** represents
  - a physical entity (record in a database), such as molecular entity, disease, clinical trial, ...
  - information artifact, such as ontological term, scientific publication, book, ...
  - ...
- **Qualifier** uses
  - term from controlled vocabulary (biomodels model/biol)

http://co.mbine.org/standards/qualifiers

URIs: standard way to encode identifiers

A Uniform Resource Identifier (URI) is a string of characters used to identify a source of information.

- http://www.ebi.ac.uk/
- http://en.wikipedia.org/wiki/Uniform_Resource_Identifier
- https://www.ebi.ac.uk/chembldb/
- ftp://ftp.uniprot.org/pub/databases/uniprot/uniref/
- urn:ietf:rfc:2396
- file:///home/username/presentation.pdf
- http://lsrn.org/taxon:4932
- ...

Why is such a system needed?

- Identifiers can be ambiguous
  - **9606**
  - "3-Fluorotoluene" in PubChem
  - "Homo sapiens" (NCBI Taxonomy); "Arabian tea" (GRIN Plant Taxonomy); bird (BOLD)
- data is often available from multiple sources on the web
  - Gene Ontology is provided by Amigo, QuickGO, BioPortal, OLS, Bio2RDF, …
- access URLs regularly change
  - May reflect institutional/infrastructure changes
- data providers begin to mint (& maintain) their own URIs…
- ***proliferation of "equivalent" URIs***

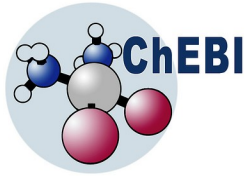
Aims and features

- Provide *unique*, *stable*, *resolvable* and *location-independent* URIs for the identification and location of Life Sciences data
- Cross-referencing & data integration
- Reliable resolving (automated checks)

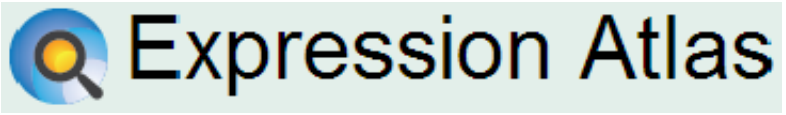
- URI scheme conversion services
- Support for different formats
- Allow data providers to maintain selected Registry records
- Enable *profiles* for customised resolving behaviour

- Curated Registry

http://identifiers.org/registry



500+ curated data collections



http://identifiers.org/registry

The Registry is a curated community driven resource, with information collated from:

- public lists (NAR, GO xref abbs, UniProt dbxref, LSRN, Bio2RDF, …)
- requests from groups and data providers
- user submissions (individual)

- Updates: 

- New submissions: tickets on SourceForge

http://sourceforge.net/p/identifiers-org/new-collection/

- Web services (SOAP, REST) and exports (XML, RDF) available

Catalogue of data collections

Data collections: *recently updated*

Recently updated | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z | Categories

Name	Namespace	Definition
BioModels Database	biomodels.db	BioModels Database is a data resource that allows biologists to store, search and retrieve published mathematical models of biological interests.
SwissLipids	swisslipid	SwissLipids is a curated resource that provides information about known lipids, including lipid structure, metabolism, interactions, and subcellular and tissue localization. Information is curated from peer-reviewed literature and referenced using established ontologies, and provided with full provenance and evidence codes for curated assertions.
GOA	goa	The GOA (Gene Ontology Annotation) project provides high-quality Gene Ontology (GO) annotations to proteins in the UniProt Knowledgebase (UniProtKB) and International Protein Index (IPI). This involves electronic annotation and the integration of high-quality manual GO annotation from all GO Consortium model organism groups and specialist groups.
Mouse Genome Database	mgd	The Mouse Genome Database (MGD) project includes data on gene characterization, nomenclature, mapping, gene homologies among mammals, sequence links, phenotypes, allelic variants and mutants, and strain data.
VectorBase	vectorbase	VectorBase is an NIAID-funded Bioinformatic Resource Center focused on invertebrate vectors of human pathogens. VectorBase annotates and curates vector genomes providing a web accessible integrated resource for the research community. Currently, VectorBase contains genome information for three mosquito species: Aedes aegypti, Anopheles gambiae and Culex quinquefasciatus, a body louse Pediculus humanus and a tick species Ixodes scapularis.
PharmGKB Disease	pharmgkb.disease	The PharmGKB database is a central repository for genetic, genomic, molecular and cellular phenotype data and clinical information about people who have participated in pharmacogenomics research studies. The data includes, but is not limited to, clinical and basic pharmacokinetic and pharmacogenomic research in the cardiovascular, pulmonary, cancer, pathways, metabolic and transporter domains.
GOLD genome	gold.genome	The GOLD (Genomes OnLine Database)is a resource for centralized monitoring of genome and metagenome projects worldwide. It stores information on complete and ongoing projects, along with their associated metadata. This collection references the sequencing status of individual genomes.
		Phenol-Explorer is an electronic database on polyphenol content in foods. Polyphenols form a wide group of natural antioxidants

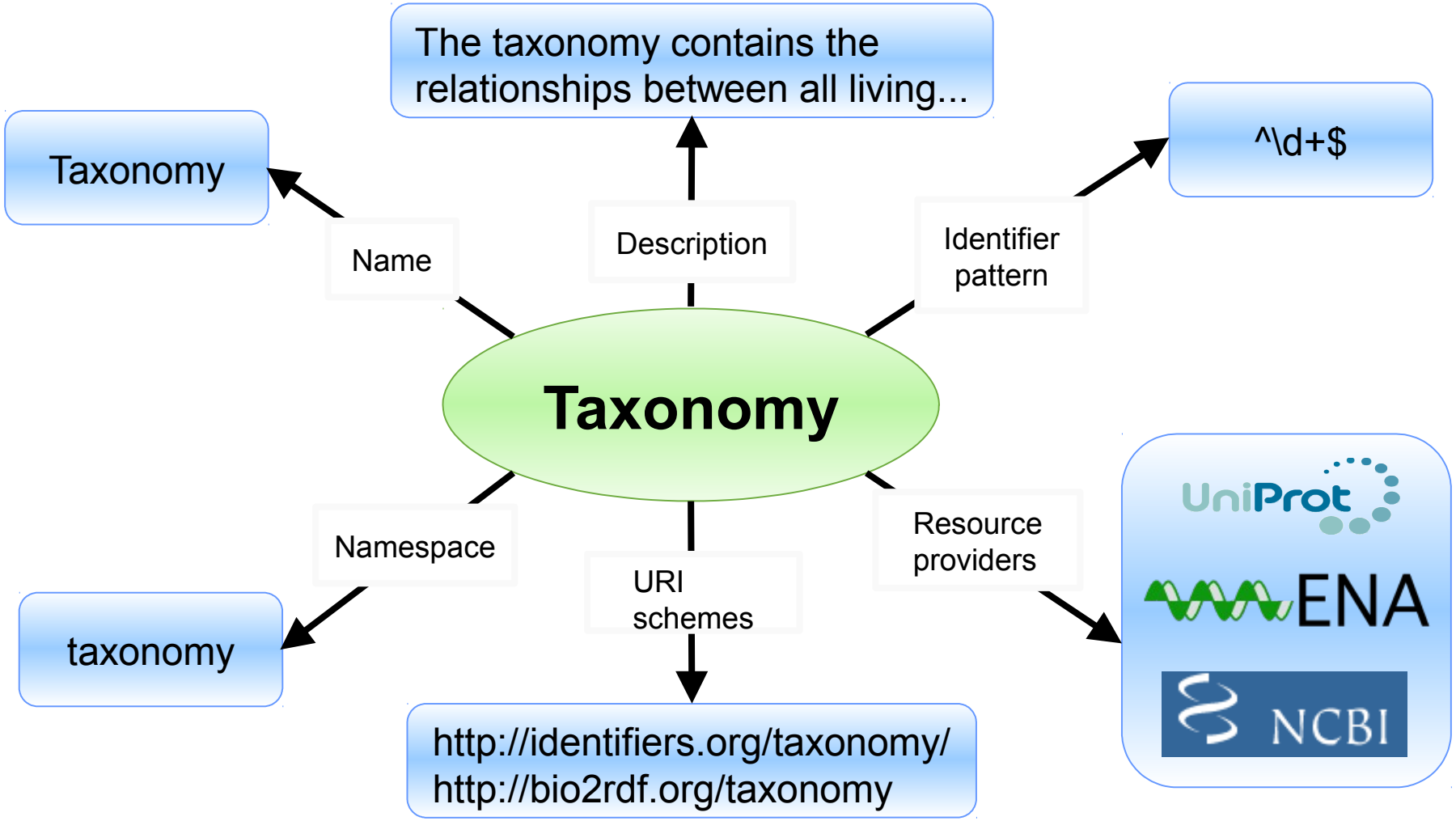
Catalogue of data collections

Data collections: *recently updated*

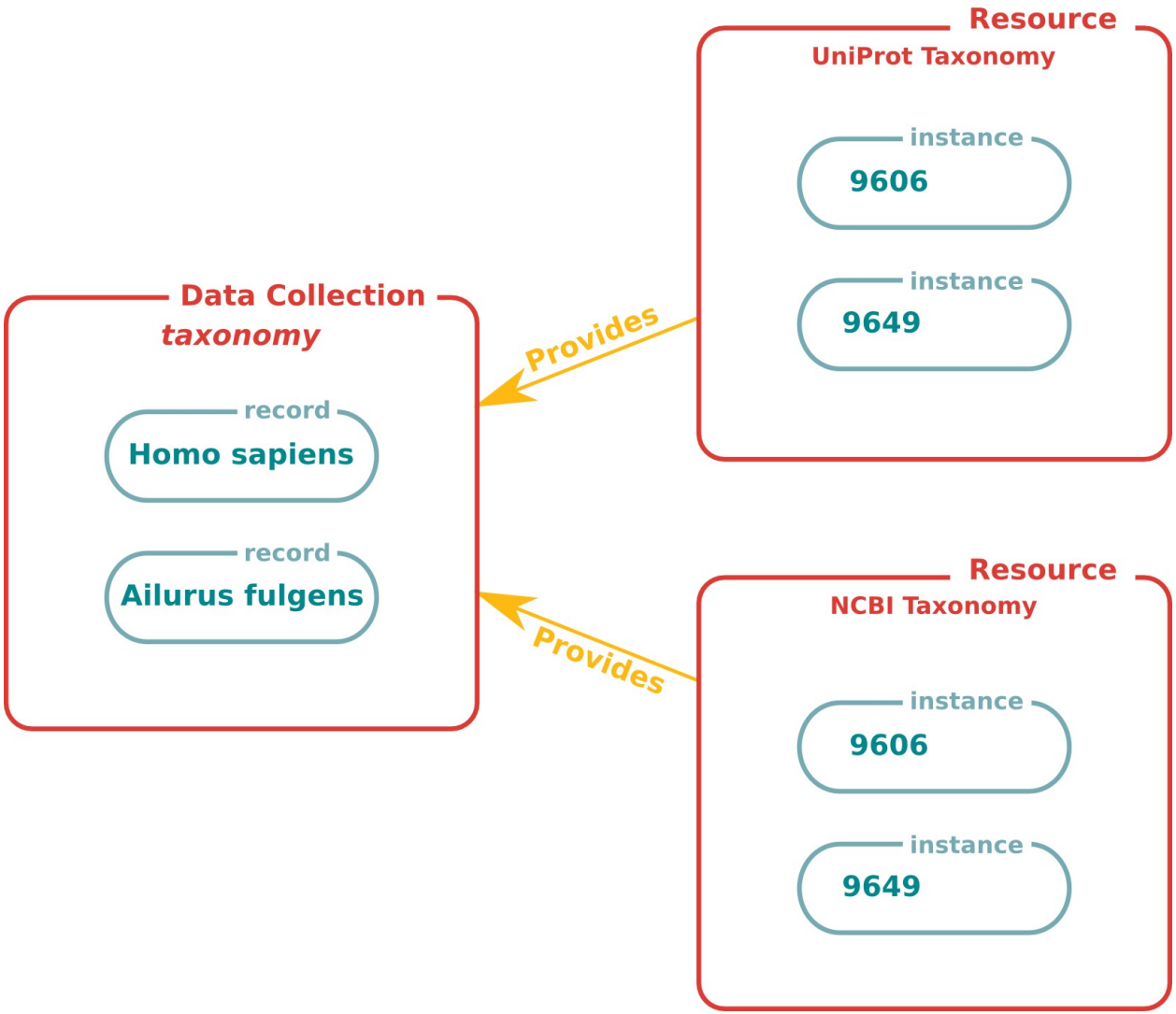
Recently updated | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z | Categories

Name	Namespace	Definition
BioModels Database	biomodels.db	BioModels Database is a data resource that allows biologists to store, search and retrieve published mathematical models of biological interests.
SwissLipids	swisslipid	SwissLipids is a curated resource that provides information about known lipids, including lipid structure, metabolism, interactions, and subcellular and tissue localization. Information is curated from peer-reviewed literature and referenced using established ontologies, and provided with full provenance and evidence codes for curated assertions.
GOA	goa	The GOA (Gene Ontology Annotation) project provides high-quality Gene Ontology (GO) annotations to proteins in the UniProt Knowledgebase (UniProtKB) and International Protein Index (IPI). This involves electronic annotation and the integration of high-quality manual GO annotation from all GO Consortium model organism groups and specialist groups.
Mouse Genome Database	mgd	The Mouse Genome Database (MGD) project includes data on gene characterization, nomenclature, mapping, gene homologies among mammals, sequence links, phenotypes, allelic variants and mutants, and strain data.
VectorBase	vectorbase	VectorBase is an NIAID-funded Bioinformatic Resource Center focused on invertebrate vectors of human pathogens. VectorBase annotates and curates vector genomes providing a web accessible integrated resource for the research community. Currently, VectorBase contains genome information for three mosquito species: Aedes aegypti, Anopheles gambiae and Culex quinquefasciatus, a body louse Pediculus humanus and a tick species Ixodes scapularis.
PharmGKB Disease	pharmgkb.disease	The PharmGKB database is a central repository for genetic, genomic, molecular and cellular phenotype data and clinical information about people who have participated in pharmacogenomics research studies. The data includes, but is not limited to, clinical and basic pharmacokinetic and pharmacogenomic research in the cardiovascular, pulmonary, cancer, pathways, metabolic and transporter domains.
GOLD genome	gold.genome	The GOLD (Genomes OnLine Database)is a resource for centralized monitoring of genome and metagenome projects worldwide. It stores information on complete and ongoing projects, along with their associated metadata. This collection references the sequencing status of individual genomes.
		Phenol-Explorer is an electronic database on polyphenol content in foods. Polyphenols form a wide group of natural antioxidants

Collection


The taxonomy contains the relationships between all living...
Taxonomy
Name
Description
Identifier pattern
^\d+$
Taxonomy
Namespace
URI schemes
Resource providers
taxonomy
http://identifiers.org/taxonomy/
http://bio2rdf.org/taxonomy
UniProt
ENA
NCBI

Decoupling data identification and access



Collection resources

Data collection: *Enzyme Nomenclature*

Overview | Miscellaneous | RDF/XML | Turtle

General information

Recommended name	**Enzyme Nomenclature**
Alternative name(s)	Enzyme Classification EC code EC
Description	The Enzyme Classification contains the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzyme-catalysed reactions.
Identifier pattern	^\d+\.-\.-\.-|\d+\.\d+\.-\.-|\d+\.\d+\.\d+\.-|\d+\.\d+\.\d+\.(n)?\d+$
Registry identifier	MIR:00000004

protein | enzyme | classification | taxonomy

Identification schemes

Namespace	ec-code
URI	http://identifiers.org/ec-code/

Alternative URI schemes

- urn:miriam:ec-code

Deprecated URI scheme(s) ⊞

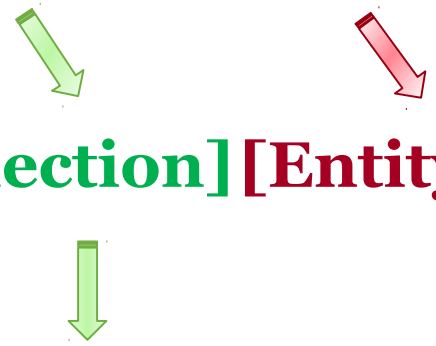
Physical locations (resources)

Resource MIR:00100308	Description: ExploreEnz at Trinity College Access URLs: HTML (using the example identifier: 1.1.1.1) Institution: Trinity College, Dublin, Ireland Website: http://www.enzyme-database.org/
Resource MIR:00100002	Description: KEGG Ligand Database for Enzyme Nomenclature Access URLs: HTML (using the example identifier: 1.1.1.1) Institution: Kyoto University Bioinformatics Center, Japan Website: http://www.genome.jp/dbget-bin/www_bfind?enzyme
Resource MIR:00100003	Description: Enzyme nomenclature database, ExPASy (Expert Protein Analysis System) Access URLs: HTML (using the example identifier: 1.1.1.1) Institution: Swiss Institute of Bioinformatics, Switzerland Website: http://enzyme.expasy.org/
Resource MIR:00100001	Description: IntEnz (Integrated relational Enzyme database) Access URLs: HTML (using the example identifier: 1.1.1.1) Institution: European Bioinformatics Institute, United Kingdom Website: http://www.ebi.ac.uk/intenz/

resources

Unique URI construction

Homo sapiens in **Taxonomy** (**9606**)



[Data collection] **[Entity identifier]**

http://identifiers.org/taxonomy/9606

Unique URI construction

- URI to identify the entity 'Homo sapiens' in the data collection Taxonomy

http://identifiers.org/taxonomy/9606

Unique URI construction

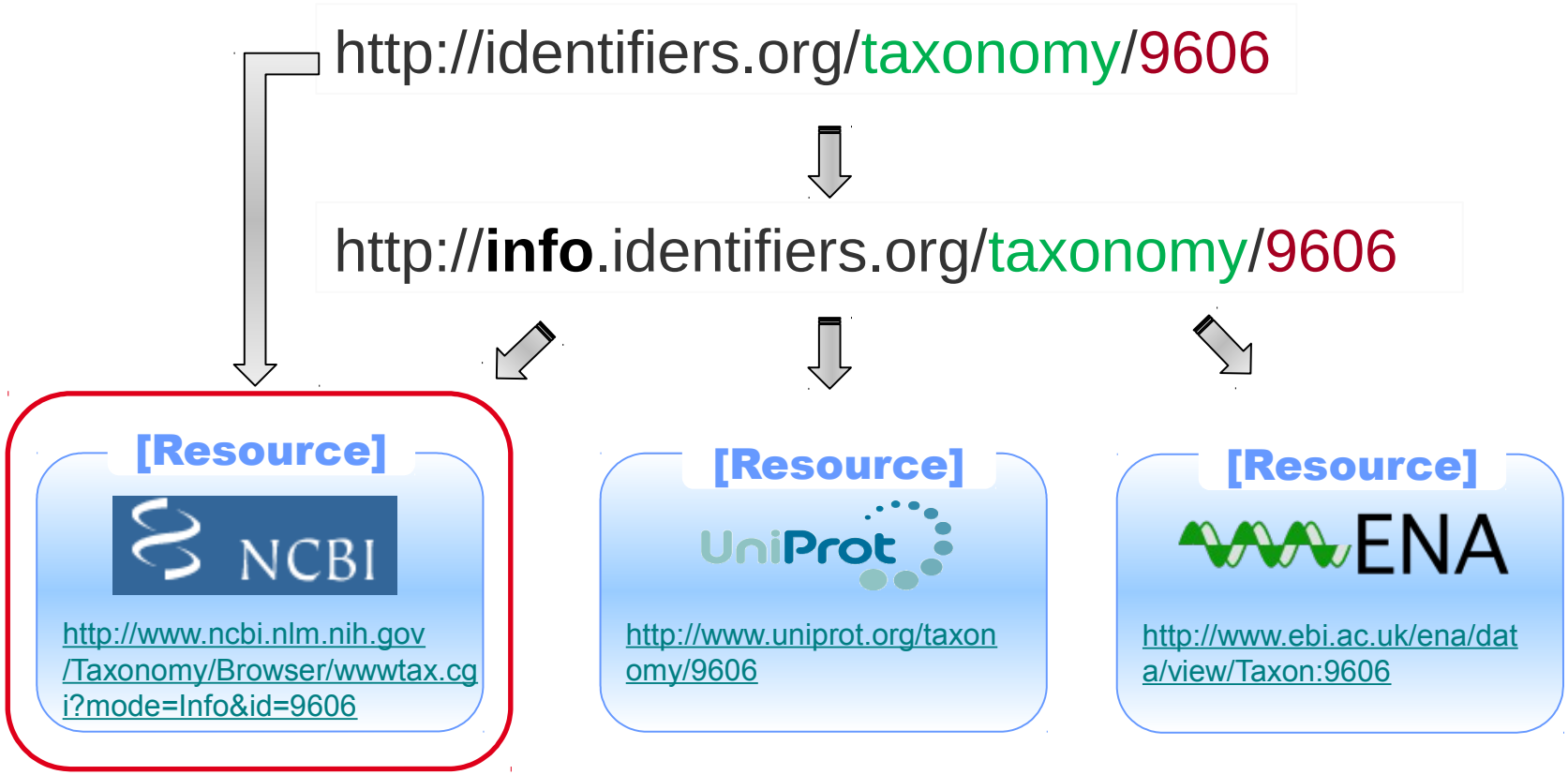
- URI to identify the Registry's record of this concept

http://identifiers.org/taxonomy/9606

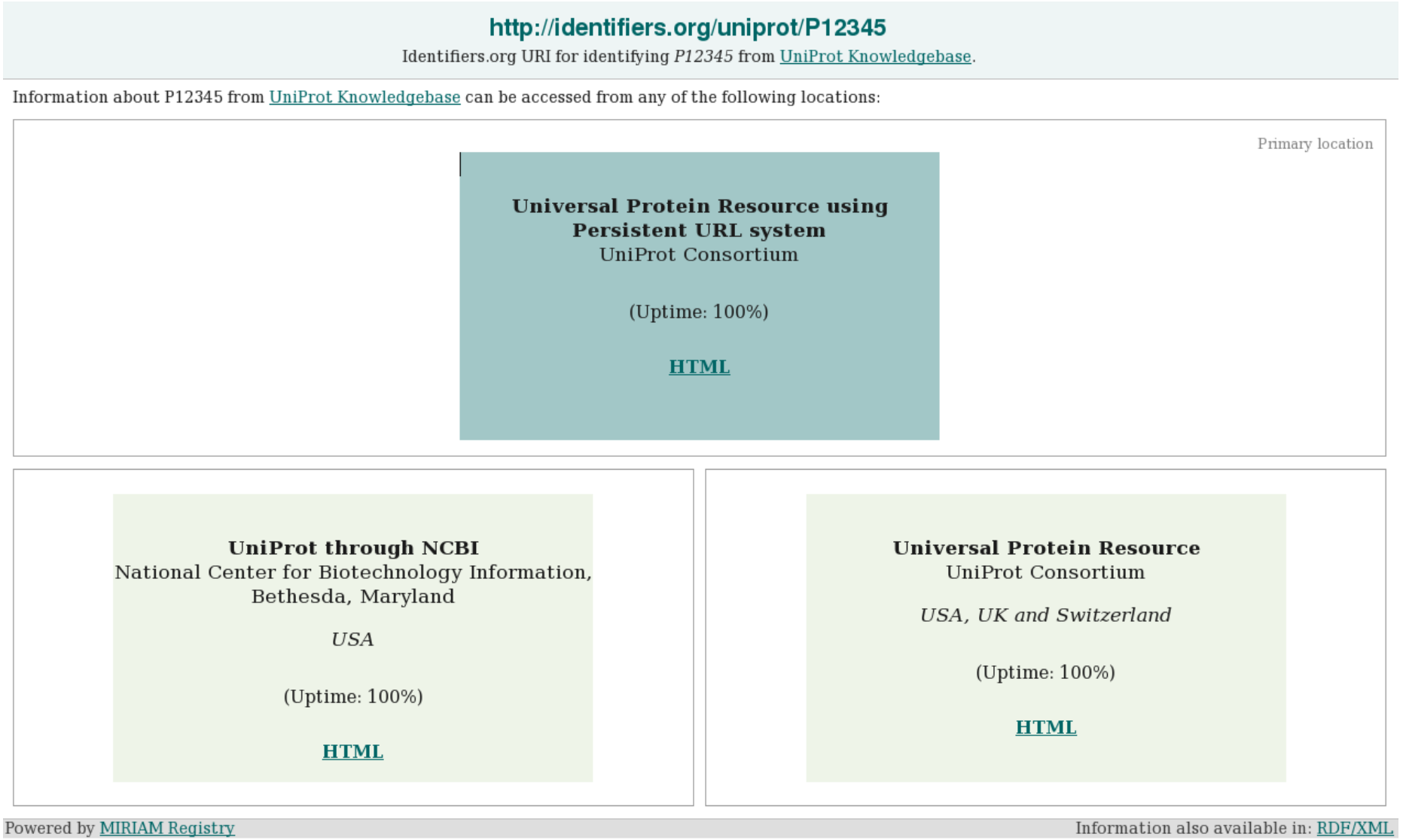
http://**info**.identifiers.org/taxonomy/9606

Unique URI construction

- URI to identify the Registry's record of this concept



Resolvable URIs: http://info.identifiers.org/uniprot/P12345

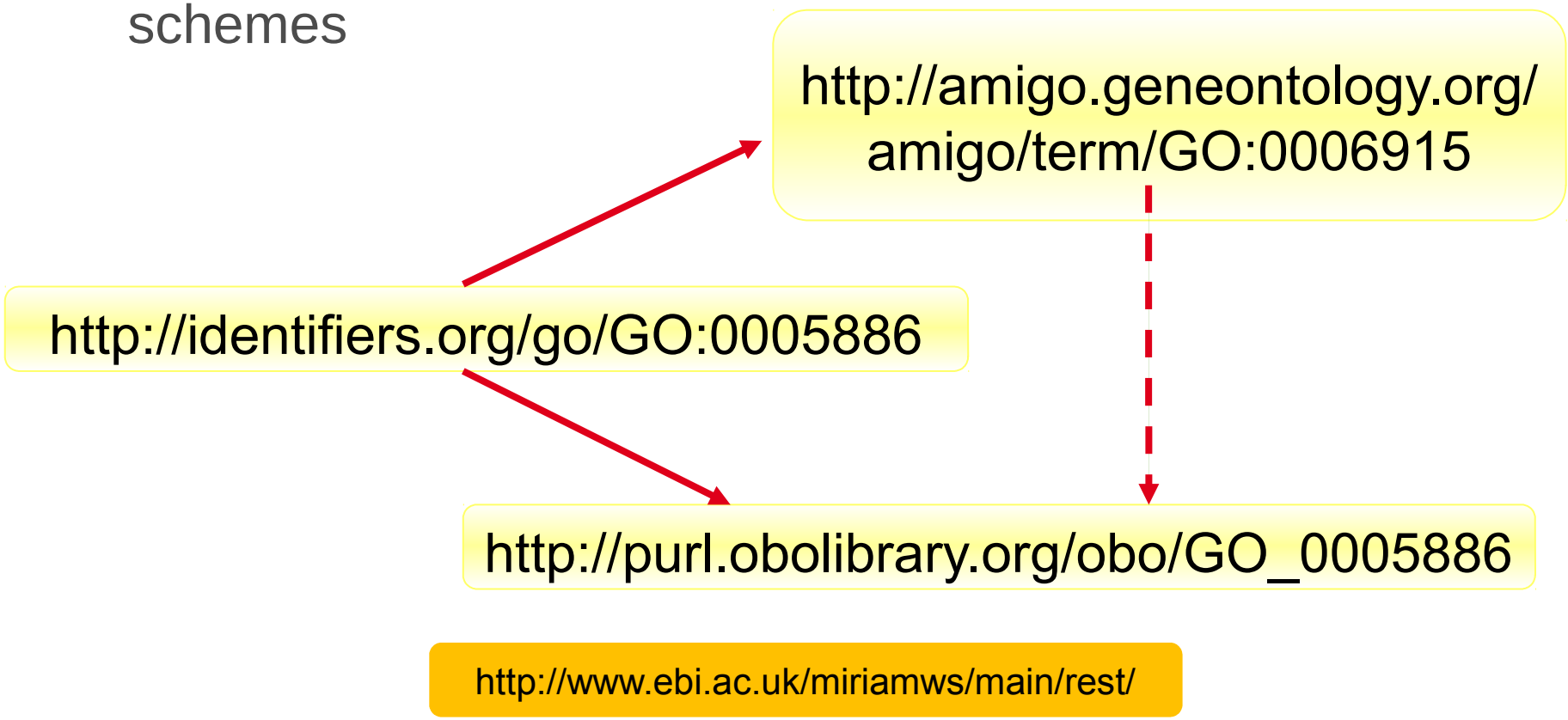


Current work: update

- URI interconversion system
- Implement profile system to allow users/groups/communities to customise resolving behaviour

Interconversion of identifier schemes

- Registry records different identifier schemes
- Web service for inter-conversion between identifier schemes



SPARQL-enabled identifier conversion

- A SPARQL-based service to performs URI scheme conversion so that users do not need to worry about the type of URIs which are being used in each dataset.

http://purl.obolibrary.org/obo/GO_0006915

Bridges the gap between the Bio2RDF-specified URIs and URIs used within the EBI RDF platform.

identifiers.org

BIO2RDF

http://bio2rdf.org/go:0006915

http://identifiers.org/services/sparql

Profiles

- Users will be able to login to the Registry

- Create profiles
  - select the data collections they use
  - preselect a preferred resolving location (or resource) for each of them

  - http://identifiers.org/pubmed/22140103?profile=[your profile name]

Profiles

Profile: biomodels

BioModels Database

BioModels Database is a repository of peer-reviewed, published, computational models. These mathematical models are primarily from the field of systems biology, but more generally are those of biological interest. This resource allows biologists to store, search and retrieve published mathematical models. In addition, models in the database can be used to generate sub-models, can be simulated online, and can be converted between different representational formats. This resource also features programmatic access via Web Services.

biomodels

- Access: private
- Nb data collections: 31
- Contact: *laibe@ebi.ac.uk*
- Created: 2011-06-07 14:20:07
- Modified: 2011-06-17 16:46:54

Key (URL)

REAS5BbW4wzsvfYSMXr2XUBCtEF%2FiqS8ONgi2B5yfbXdu5HJ2l5p2dVBRq%2FRG1tY [Update key]

Links

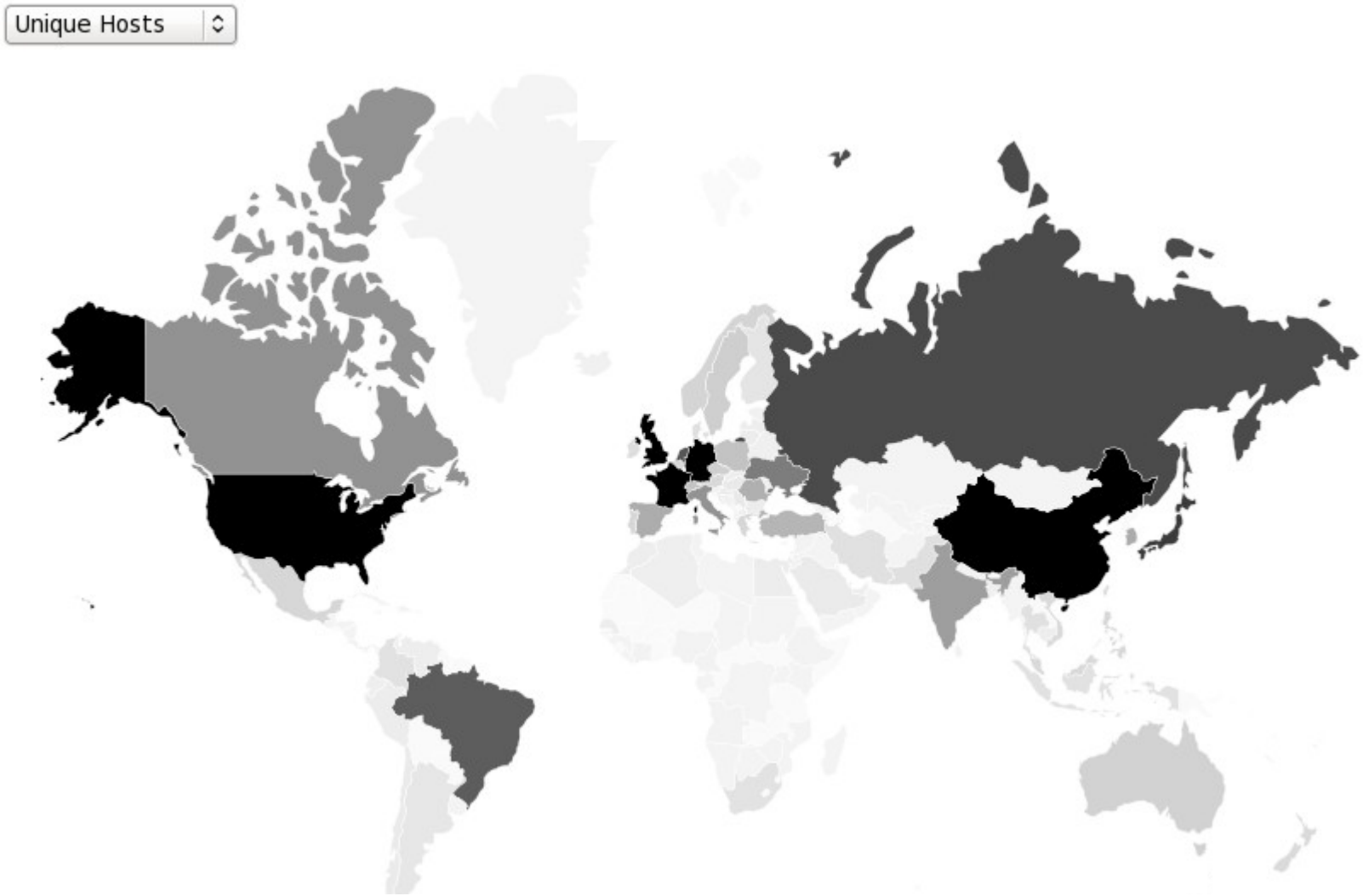
profile specific XML export

Data collections

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z | Selected

Selected	Name	Resource(s)	Date added	Date modified
☑	BioModels Database	MIR:00100006 - BioModels Database	2011-06-07 14:29:10	2011-06-17 16:39:49
☑	PubMed	MIR:00100032 - CiteXplore [OBSOLETE]	2011-06-17 16:43:14	2011-06-17 16:43:14
☑	Brenda Tissue Ontology	MIR:00100144 - Brenda Tissue Ontology through OLS	2011-06-17 16:15:13	2011-06-17 16:15:13
☑	Cell Type Ontology	MIR:00100143 - Cell Type Ontology through OLS	2011-06-17 16:18:57	2011-06-17 16:18:57
☑	CluSTr	MIR:00100030 - CluSTr Database [OBSOLETE]	2011-06-17 16:18:57	2011-06-17 16:18:57
☑	ChEBI	MIR:00100009 - ChEBI (Chemical Entities of Biological Interest)	2011-06-17 16:18:57	2011-06-17 16:18:57
☑	DOI	MIR:00100010 - Digital Object Identifier	2011-06-17 16:19:27	2011-06-17 16:19:27
☑	Enzyme Nomenclature	MIR:00100001 - IntEnZ (Integrated relational Enzyme database)	2011-06-17 16:19:58	2011-06-17 16:19:58
☑	NCBI Gene	MIR:00100099 - Entrez Gene (NCBI)	2011-06-17 16:19:58	2011-06-17 16:19:58

Identifiers.org web stats



Semantic Web

Common framework allowing **data** to be **shared** and **reused** across applications.

Goal: allows **machines** to understand the **semantics**, or meaning, of information on the **World Wide Web**.

Collaborative effort led by **W3C** with participation from a large number of researchers and industrial partners.

Information as triples

Subject	Predicate	Object
EGFR membrane	*located_in*	plasma
P00533 (UniProt)	OBO_REL:0000008 (OBO Relation ontology)	GO:0005886 (Gene Ontology)

http://identifiers.org/uniprot/P00533

http://identifiers.org/obo.ro/OBO_REL:0000008

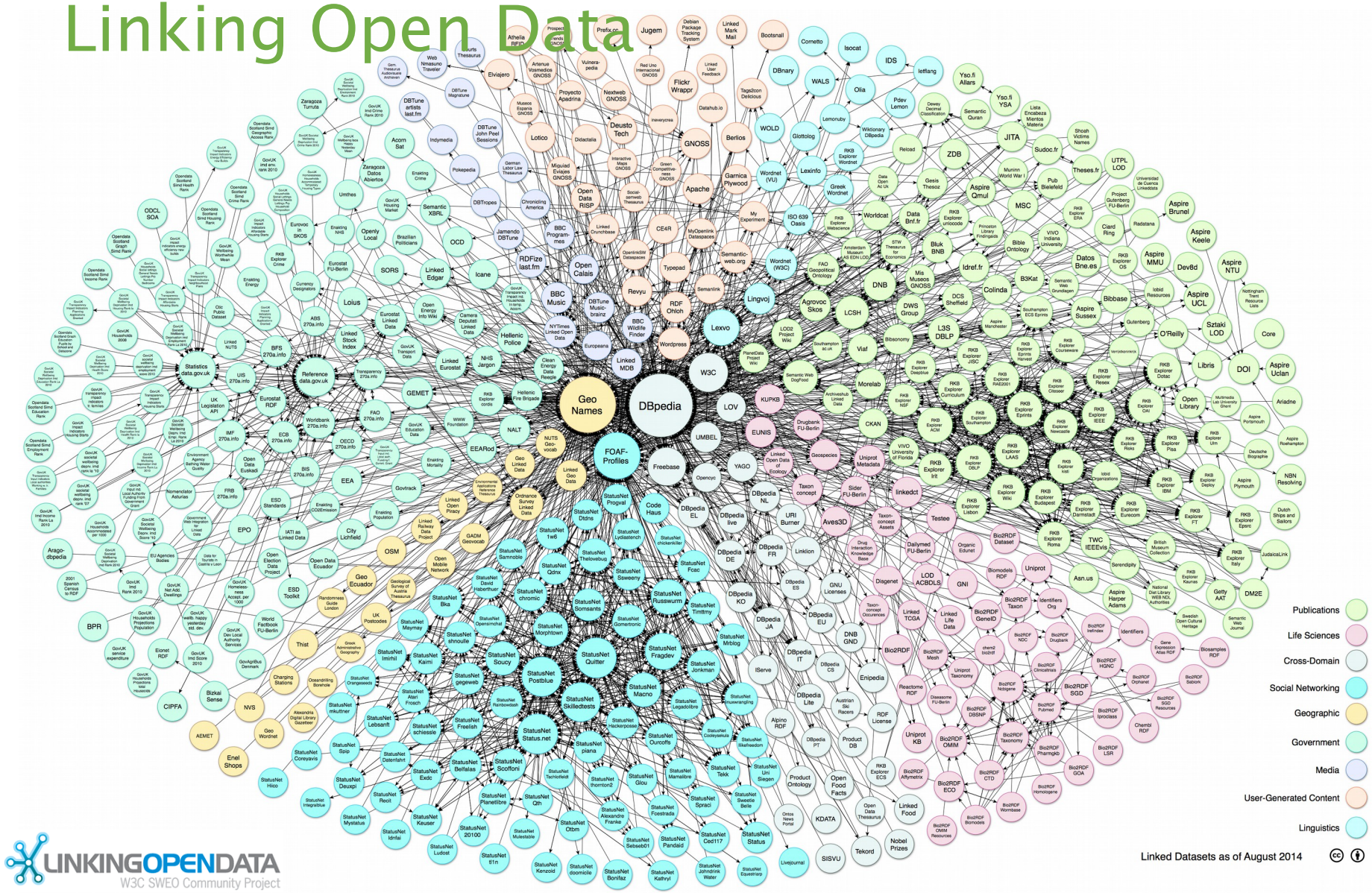
http://identifiers.org/go/GO:0005886

Semantic Web: technologies

Set of technologies helping to link data of different natures and different locations in a meaningful way:

- **Uniform Resource Identifiers** (URIs) to unambiguously identify pieces of data
- **Controlled vocabularies** to characterise the relationships between data entities (SKOS, RDFs, ontologies)
- **Syntaxes** to encode the relationships between data entities (RDF)
- **Query** languages, to retrieve information encoded using semantic web technologies (SPARQL)

Linking Open Data



LINKINGOPENDATA
W3C SWEO Community Project

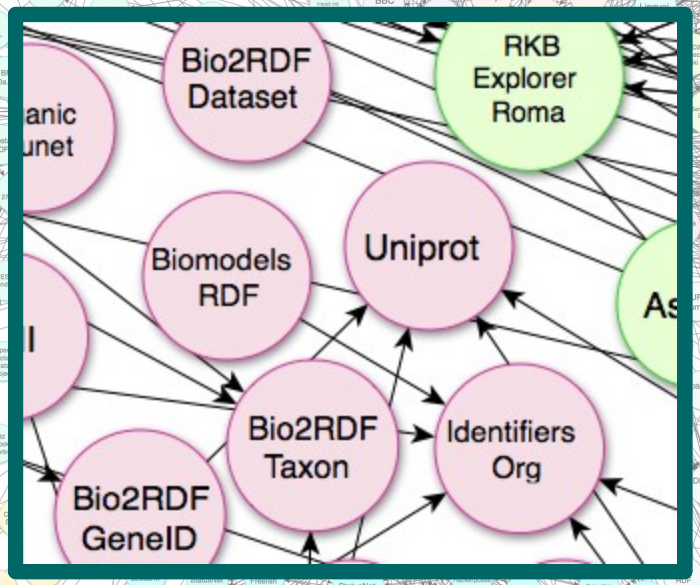
COMBINE 2015
Salt Lake City
Utah

identifiers.org

EMBL-EBI

Linking Open Data

http://lod-cloud.net/



Publications
Life Sciences
Cross-Domain
Social Networking
Geographic
Government
Media
User-Generated Content
Linguistics

Linked Datasets as of August 2014

LINKINGOPENDATA
W3C SWEO Community Project

How to contribute

- Identifiers.org : http://identifiers.org/
- Registry : http://identifiers.org/registry
- Mailing lists
  - Support : biomodels-net-support@lists.sf.net
  - Discussion : IdOrg-discuss@googlegroups.com
- Requests : https://sourceforge.net/projects/identifiers-org/

We appreciate your support!

Acknowledgements

- Computational systems biology community
- BioModels.net team
- Nicolas Le Novère
- Camille Laibe

juty@ebi.ac.uk

biomodels-net-support@lists.sf.net

Summary

- free to use
- stable
- resolvable
- customisable

"perennial URIs for data integration and cross-referencing"

Homo sapiens in Taxonomy (**9606**)

[data collection] [entity identifier]

http://identifiers.org/taxonomy/**9606**

URIs for the user

- where to get the data
- what to use to cross-reference data
- how to share with your colleagues

URIs for data providers

- store in database
- use in import/export formats
- display in user interfaces

Services: curated Registry, virtual SPARQL endpoint, XML and RDF exports, web services, ...

Annotations in SBML

```
[...]
<species metaid="metaid_0000006"
         id="L_EGFR"
         compartment="compartment"
         initialConcentration="0">
  <annotation>
    <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22rdfsyntaxns#"
             xmlns:bqbiol="http://biomodels.net/biologyqualifiers/">
      <rdf:Description rdf:about="#metaid_0000006">
        <bqbiol:hasPart>
          <rdf:Bag>
            <rdf:li rdf:resource="http://identifiers.org/uniprot/P07522" />
            <rdf:li rdf:resource="http://identifiers.org/uniprot/Q9QX70" />
          </rdf:Bag>
        </bqbiol:hasPart>
      </rdf:Description>
    </rdf:RDF>
  </annotation>
</species>
[...]
```