identifiers.org

Integration tool for heterogeneous datasets

Sarala M Wimalaratne[1], Nick Juty[1], Nicolas Le Novère[1,2], Henning Hermjakob[1] and Camille Laibe[1]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD
[2]Babraham Institute, Babraham Research Campus, Cambridge CB22 3AT, http://www.babraham.ac.uk/

EMBL-EBI

http://identifiers.org

Introduction

Data providers identify individual records within their datasets using an identifier or accession number. These identifiers are only unique within the dataset they originate from. Additionally, the data is often distributed via multiple resources, using different International Resource Identifiers (IRI) to access individual records.

Identifiers.org [1] provides resolvable persistent IRIs used to identify individual records (based on the existing entity identifiers assigned directly by the data providers). IRI assignment is reliant on an underlying Registry which contains information about individual data collections, the resources (or physical locations) which distribute them, and the different root IRIs used to identify and access the individual records.

Creating Unique IRIs

Canonical Identifiers.org IRIs for uniquely identifying data entities:

http://identifiers.org/[collection]/[entity]

[collection] : namespace of a data collection

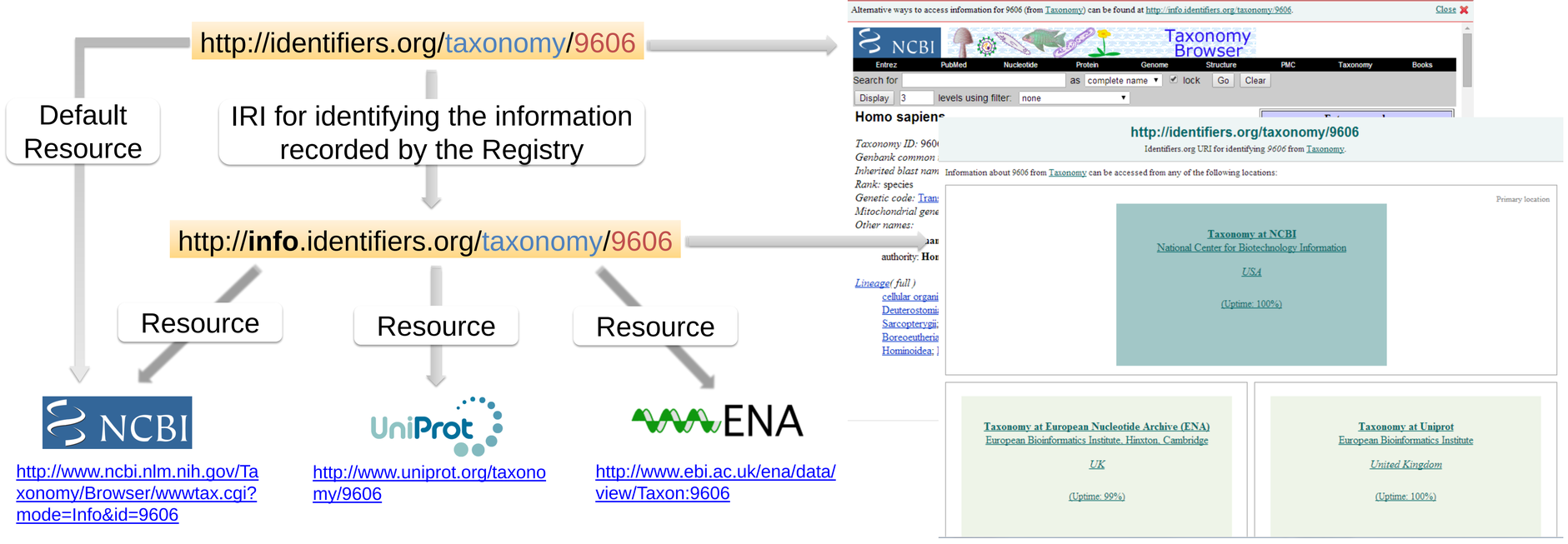
[entity] : identifier of the entity created by the original data provider

Homo sapiens in Taxonomy (9606)

[collection] [entity]

http://identifiers.org/taxonomy/9606

Resolving Service



When resolving the canonical IRI, a default resource is used to provide a preview of the identified entity. Its selection uses an algorithm which relies on various properties, such as whether the resource is the primary provider for the content and its reliability (uptime).

Registry

The Registry is a catalogue of data collections, which can be ontologies or primary data resources available on the Web. Each of these is uniquely identified within the Registry and assigned a namespace (short unique string of characters identifying the data collection within the canonical Identifiers.org IRIs). Information is also stored for all resources through which collection records can be accessed.

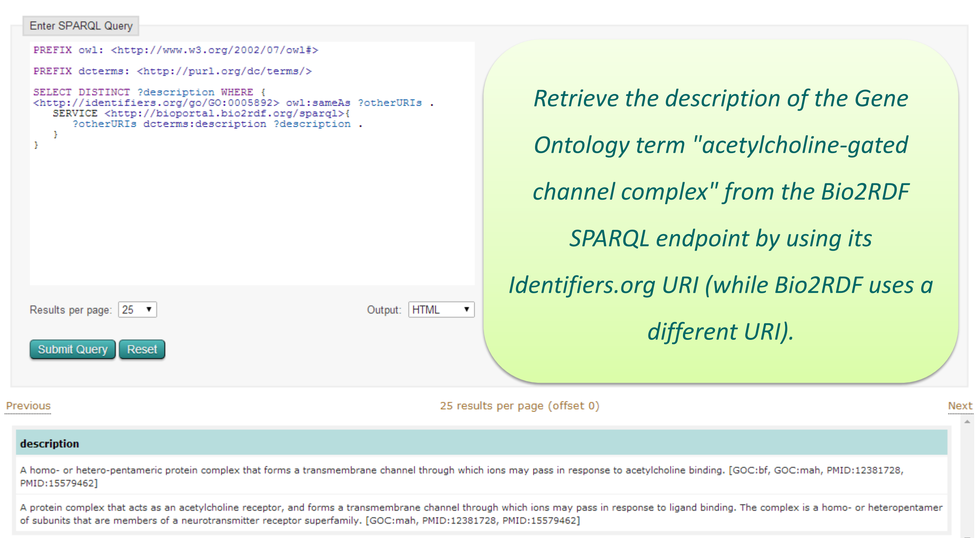
Over 500 curated Data collections

http://identifiers.org/registry

Virtual SPARQL Endpoint

We provide a “virtual” SPARQL endpoint, directly driven by the database underpinning the Registry, and thus is not based on an actual triple store. It uses Registry-stored information about IRI patterns to automatically translate identifiers from one IRI scheme to another. Specifically, the service has been designed to answer queries requesting IRIs and makes use of the predicate owl:sameAs.

This service provides an automatic system for linking and querying the ever growing number of life science data resources such as EBI RDF Platform [2] and Bio2RDF [3].



http://identifiers.org/services/sparql

Get Involved

We welcome any and all user contributions.

- Report issues related to the services or content: **biomodels-net-support@lists.sf.net**
- General discussion about the resource: **IdOrg-discuss@googlegroups.com**
- Request the addition of new data collection(s): **http//sourceforge.net/p/identifiers-org/new-collection/**

References

1. Juty N, Lè Novère N, Laibe C: **Identifiers.org and MIRIAM Registry: community resources to provide persistent identification.** *Nucleic Acids Res* 2011, 40(iv):1–7
2. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J: **Bio2RDF: Towards a mashup to build bioinformatics knowledge systems.** *J Biomed Inform* 2008, 41:706–716.
3. Jupp S, Malone J, Bolleman J, Brandizi M, Davies M, Garcia L, Gaulton A, Gehant S, Laibe C, Redaschi N, Wimalaratne SM, Martin M, Le Novere N, Parkinson H, Birney E, Jenkinson AM: **The EBI RDF Platform: Linked Open Data for the Life Sciences.** *Bioinformatics* 2014:btt765

Funders

EMBL-EBI, Babraham Institute, BBSRC bioscience for the future, Open PHACTS, elixir

EMBL-EBI
Wellcome Trust Genome Campus
Hinxton, Cambridgeshire, CB10 1SD, UK

Tel. +44 (0) 1223 494 521
sarala@ebi.ac.uk
www.ebi.ac.uk