Identifiers.org: Practical integration tool for heterogeneous datasets

Nick Juty

Outline

- Introduction
  - Registry
  - Identifiers.org
- Current status
- Future plans
- Invitation to contribute

Objective

provision of a multi-purpose cross-referencing and integration system

Unique stable, resolvable, location-independent URI

Community driven

Submission open

Free to use

- Curated registry

Catalogue of collections ..

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
Ensembl	ensembl	Ensembl is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. This collections also references outgroup organisms.
RefSeq	refseq	The Reference Sequence (RefSeq) collection aims to provide a comprehensive, integrated, non-redundant set of sequences, including genomic DNA, transcript (RNA), and protein products.
National Drug Code	ndc	The National Drug Code (NDC) is a unique, three-segment number used by the Food and Drug Administration (FDA) to identify drug products for commercial use. This is required by the Drug Listing Act of 1972. The FDA publishes and updates the listed NDC numbers daily.
ClinicalTrials.gov	clinicaltrials	ClinicalTrials.gov provides free access to information on clinical studies for a wide range of diseases and conditions. Studies listed in the database are conducted in 175 countries
IMGT HLA	imgt.hla	IMGT, the international ImMunoGeneTics project, is a collection of high-quality integrated databases specialising in Immunoglobulins, T cell receptors and the Major Histocompatibility Complex (MHC) of all vertebrate species. IMGT/HLA is a database for sequences of the human MHC, referred to as HLA. It includes all the official sequences for the WHO Nomenclature Committee For Factors of the HLA System. This collection references allele information through the WHO nomenclature.
Gene Ontology	go	The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism.
CYGD	cygd	The MIPS Comprehensive Yeast Genome Database (CYGD) provides information on the molecular structure and functional network of the entirely sequenced the budding yeast, Saccharomyces cerevisiae, as well as on related yeasts which are used for comparative analysis.
RNA Modification Database	rnamods	The RNA modification database provides a comprehensive listing of post-transcriptionally modified nucleosides from RNA. The database consists of all RNA-derived ribonucleosides of known structure, including those from established sequence positions, as well as those detected or characterized from hydrolysates of RNA.
TreeFam	treefam	TreeFam is a database of phylogenetic trees of gene families found in animals. Automatically generated trees are curated, to create a curated resource that presents the accurate evolutionary history of all animal gene families, as

.. and their namespaces

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
Ensembl	ensembl	Ensembl is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. This collections also references outgroup organisms.
RefSeq	refseq	The Reference Sequence (RefSeq) collection aims to provide a comprehensive, integrated, non-redundant set of sequences, including genomic DNA, transcript (RNA), and protein products.
National Drug Code	ndc	The National Drug Code (NDC) is a unique, three-segment number used by the Food and Drug Administration (FDA) to identify drug products for commercial use. This is required by the Drug Listing Act of 1972. The FDA publishes and updates the listed NDC numbers daily.
ClinicalTrials.gov	clinicaltrials	ClinicalTrials.gov provides free access to information on clinical studies for a wide range of diseases and conditions. Studies listed in the database are conducted in 175 countries
IMGT HLA	imgt.hla	IMGT, the international ImMunoGeneTics project, is a collection of high-quality integrated databases specialising in Immunoglobulins, T cell receptors and the Major Histocompatibility Complex (MHC) of all vertebrate species. IMGT/HLA is a database for sequences of the human MHC, referred to as HLA. It includes all the official sequences for the WHO Nomenclature Committee For Factors of the HLA System. This collection references allele information through the WHO nomenclature.
Gene Ontology	go	The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism.
CYGD	cygd	The MIPS Comprehensive Yeast Genome Database (CYGD) provides information on the molecular structure and functional network of the entirely sequenced the budding yeast, Saccharomyces cerevisiae, as well as on related yeasts which are used for comparative analysis.
RNA Modification Database	rnamods	The RNA modification database provides a comprehensive listing of post-transcriptionally modified nucleosides from RNA. The database consists of all RNA-derived ribonucleosides of known structure, including those from established sequence positions, as well as those detected or characterized from hydrolysates of RNA.
TreeFam	treefam	TreeFam is a database of phylogenetic trees of gene families found in animals. Automatically generated trees are curated, to create a curated resource that presents the accurate evolutionary history of all animal gene families, as

Sources

Collection information collated from:

- **public lists** (NAR, GO, LSRN, Bio2RDF …)
- **requests** from groups (list databases used)
- **user** submissions (individual)

Data Collection

Data collection: *Enzyme Nomenclature*

Overview | Categories | Miscellaneous

Identification		
Identifier		MIR:00000004
Name		Enzyme Nomenclature
Synonyms		Enzyme Classification EC code EC
Information		
Definition		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+$
URIs		
Namespace		ec-code
Root URL		http://identifiers.org/ec-code/
Root URN		urn:miriam:ec-code:
Physical Locations		
Resource MIR:00100308	**Access URL**	http://www.enzyme-database.org/query.php?ec=$id [Example: 1.1.1.1]
	Website	http://www.enzyme-database.org/
	Description	ExploreEnz at Trinity College
	Institution	Trinity College, Dublin, Ireland
Resource MIR:00100002	**Access URL**	http://www.genome.jp/dbget-bin/www_bget?ec:$id [Example: 1.1.1.1]
	Website	http://www.genome.jp/dbget-bin/www_bfind?enzyme
	Description	KEGG Ligand Database for Enzyme Nomenclature
	Institution	Kyoto University Bioinformatics Center, Japan
Resource MIR:00100001	**Access URL**	http://www.ebi.ac.uk/intenz/query?cmd=SearchEC&ec=$id [Example: 1.1.1.1]
	Website	http://www.ebi.ac.uk/intenz/
	Description	IntEnZ (Integrated relational Enzyme database)
	Institution	European Bioinformatics Institute, United Kingdom
Resource MIR:00100003	**Access URL**	http://enzyme.expasy.org/EC/$id [Example: 1.1.1.1]
	Website	http://enzyme.expasy.org/
	Description	Enzyme nomenclature database, ExPASy (Expert Protein Analysis System)
	Institution	Swiss Institute of Bioinformatics, Switzerland
References		

..associated resolving locations

Data collection: *Enzyme Nomenclature*

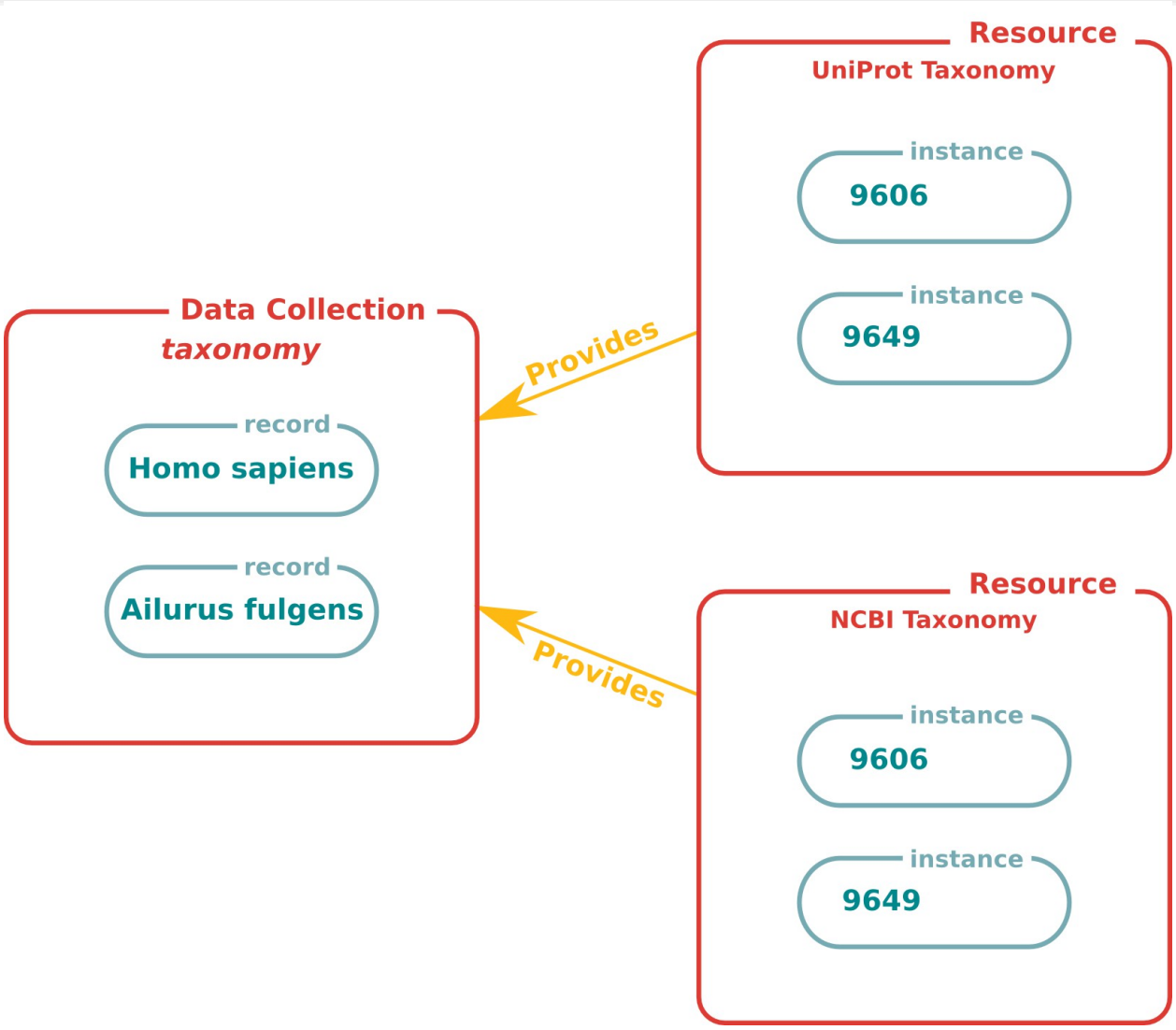
Overview | Categories | Miscellaneous

Identification	
Identifier	MIR:00000004
Name	Enzyme Nomenclature
Synonyms	Enzyme Classification EC code EC
Information	
Definition	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+$
URIs	
Namespace	ec-code
Root URL	http://identifiers.org/ec-code/
Root URN	urn:miriam:ec-code:

Physical Locations		
Resource MIR:00100308	**Access URL**	http://www.enzyme-database.org/query.php?ec=$id [Example: 1.1.1.1]
	Website	http://www.enzyme-database.org/
	Description	ExploreEnz at Trinity College
	Institution	Trinity College, Dublin, Ireland
Resource MIR:00100002	**Access URL**	http://www.genome.jp/dbget-bin/www_bget?ec:$id [Example: 1.1.1.1]
	Website	http://www.genome.jp/dbget-bin/www_bfind?enzyme
	Description	KEGG Ligand Database for Enzyme Nomenclature
	Institution	Kyoto University Bioinformatics Center, Japan
Resource MIR:00100001	**Access URL**	http://www.ebi.ac.uk/intenz/query?cmd=SearchEC&ec=$id [Example: 1.1.1.1]
	Website	http://www.ebi.ac.uk/intenz/
	Description	IntEnZ (Integrated relational Enzyme database)
	Institution	European Bioinformatics Institute, United Kingdom
Resource MIR:00100003	**Access URL**	http://enzyme.expasy.org/EC/$id [Example: 1.1.1.1]
	Website	http://enzyme.expasy.org/
	Description	Enzyme nomenclature database, ExPASy (Expert Protein Analysis System)
	Institution	Swiss Institute of Bioinformatics, Switzerland
References		

Resources

Decoupling data and location



- built on the information stored in the **Registry**
- provides **resolvable URIs**
- resolves to **intermediate** page (customizable)

Alcohol dehydrogenase: 1.1.1.1 in Enzyme Nomenclature

↳ `http://identifiers.org/ec-code/1.1.1.1`

Activation of MAPKK activity: GO:0000186 in Gene Ontology

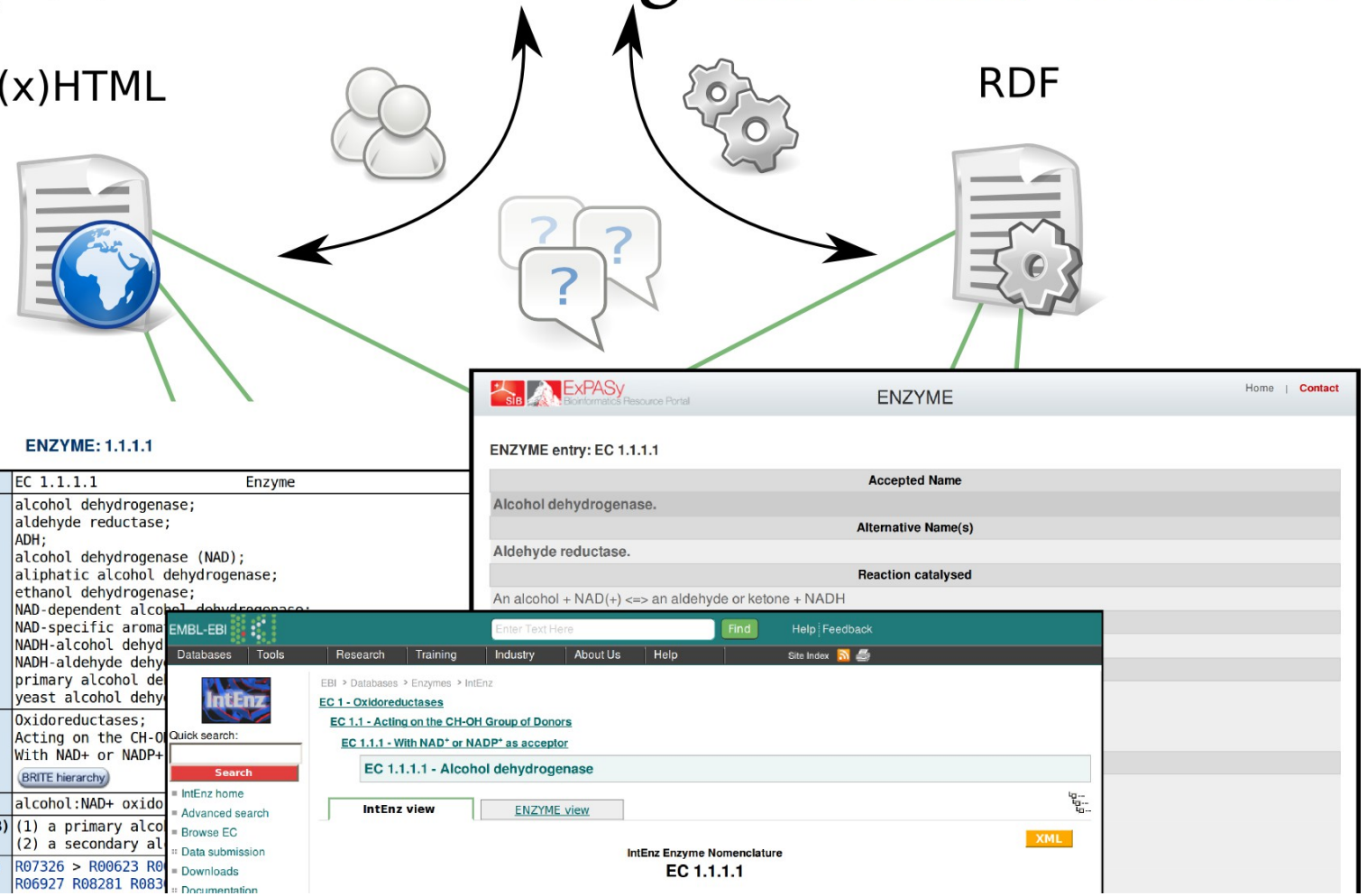
↳ `http://identifiers.org/go/GO:0000186`

http://identifiers.org/

Juty, Le Novère, Laibe. **Identifiers.org and MIRIAM Registry: community resources to provide persistent identification.** Nucleic Acids Research, 2012

Resolvable URIs

http://identifiers.org/ec-code/1.1.1.1



RDF/XML version

```
−<sio:SIO_000671>
 −<edam:data_2091>
    <sio:SIO_000300>1.1.1.1</sio:SIO_000300>
    <rdf:type rdf:resource="http://idtype.identifiers.org/ec-code/"/>
  </edam:data_2091>
 </sio:SIO_000671>

<!-- physical locations (resources) -->
−<rdfs:seeAlso>
 −<rdf:Description rdf:about="http://www.enzyme-database.org/query.php?ec=1.1.1.1">
    <dcterms:format>application/xhtml+xml</dcterms:format>
    <dcterms:publisher rdf:resource="http://identifiers.org/miriam.resource/MIR:00100308"/>
  </rdf:Description>
 </rdfs:seeAlso>
−<rdfs:seeAlso>
 −<rdf:Description rdf:about="http://enzyme.expasy.org/EC/1.1.1.1">
    <dcterms:format>application/xhtml+xml</dcterms:format>
    <dcterms:publisher rdf:resource="http://identifiers.org/miriam.resource/MIR:00100003"/>
  </rdf:Description>
 </rdfs:seeAlso>

  information about the data collection MIR:00000004
 -->
−<rdf:Description rdf:about="http://identifiers.org/miriam.collection/MIR:00000004">
    <dcterms:identifier>MIR:00000004</dcterms:identifier>
    <dcterms:title xml:lang="en-GB">Enzyme Nomenclature</dcterms:title>
    <dcterms:alternative>EC</dcterms:alternative>
    <dcterms:alternative>EC code</dcterms:alternative>
    <dcterms:alternative>Enzyme Classification</dcterms:alternative>
 </rdf:Description>

 information about Identifiers.org -->
−<rdf:Description rdf:about="http://identifiers.org/">
    <dcterms:title xml:lang="en-GB">Identifiers.org</dcterms:title>
    <dcterms:alternative xml:lang="en-GB">MIRIAM Resolver</dcterms:alternative>
 −<dcterms:description>
    General cross-referencing and annotation framework, providing unique and perennial URIs for life science data.
  </dcterms:description>
 −<rdfs:seeAlso>
  −<rdf:Description rdf:about="http://www.ebi.ac.uk/miriam/">
```

News

Registry redesign

- new EBI guidelines
- simplified interface

- BBSRC funded position
- dedicated developer Identifiers.org
- starting summer 2013

#1 Interconversion of URIs

Current

store access URL information for each data collection
updated as required (previous 'lost')
no alternative URI forms (identification schemes)

Planned

store 'legacy' access URL information (changes on resource)
store information for alternative identification schemes (collection)

- provide the capability to generate Identifiers.org URIs from current/legacy access URLs
- provide the ability to transition from alternative identification schemes to Identifiers.org URIs
- provide a means to do the reverse process

#2 SPARQL compliant service

Follows from #1

Situation: different providers use different URIs (from varying identification schemes: OBO, Bio2RDF, …) in their public-facing triple stores

Motivation: interrelate these different URIs, regardless of their origin, within a federated SPARQL query

- enable URI mapping through SPARQL-compatible endpoint

#3 Format availability

Situation: data from providers available in multiple formats (HTML, RDF, JSON, Fasta, …)
different access URLs (resolving location, URL construction)

Motivation: provide access to different formats of the records, as required by the user

- catalogue the different formats at the level of the resource
- provide direct access (profile/parameter/content negotiation)

#4 Community involvement

Situation: increased use of Identifiers.org means there is more data to capture

Motivation: move towards community collaborative methods

Involve data providers early

- promote ownership by providers
- maintenance by provider
- curator moderation

Requires

- updated software infrastructure and user interface
- user accounts handling
- user training

How to contribute

- We invite new ideas for services

  biomodels-net-support@lists.sf.net

- We encourage new submission and modifications

  https://sourceforge.net/projects/identifiers-org/

- We appreciate your support

Acknowledgments

Trustees

Michel Dumontier
Michael Galperin
Pascale Gaudet
Lee Harland
Henning Hermjakob
Michael Hucka
Toshiaki Katayama
Nicolas Le Novère
Philippe Rocca-Serra
Mark Wilkinson

Team

Camille Laibe
Nicolas Le Novère
Henning Hermjakob
Nick Juty

Contact

juty@ebi.ac.uk