



# Assign locally, resolve globally: prefix-based collection access

Tim Clark, Nick Juty, John Kunze

*PIDapalooza*  
9th-10<sup>th</sup> Nov., 2016  
Reykjavik, Iceland



# Outline

- Background – the Data Citation Use Case
- Compact Identifiers for Biomedical Data
- Identifiers.org
- n2t.net & EZID
- Prefix alignment
  - procedure
  - examples
  - next steps

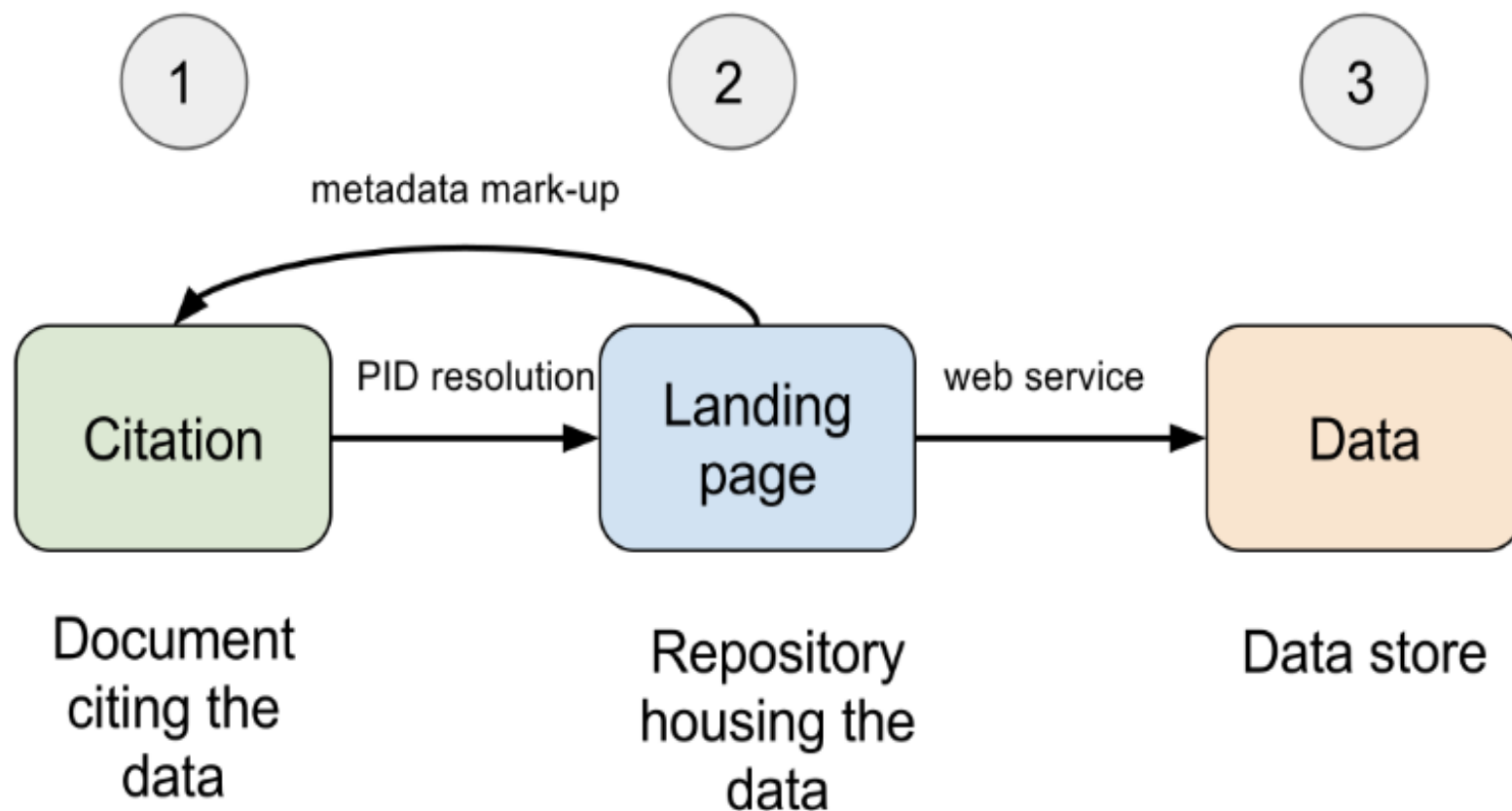
# The Data Citation Use Case

- Science policy bodies concerned about data
  - e.g. NIH, US National Academies, CODATA, Royal Society, JISC, ELIXIR
- Authorities recommend archiving / citation of primary data for reproducibility and reuse
  - Joint Declaration of Data Citation Principles (JDDCP 2014) defines approach
- NIH bioCADDIE Data Citation Pilot (DCIP):
  - Publishers, repositories, identifier / metadata registries implement in practice

# Citing biomedical data

- Transparency and validation: better science
  - 75-90% academic studies non-reproducible in drug development!
- Big data meta-analysis: reuse & discovery
- Radically improve translation of research into effective cures for devastating diseases
  - Cancers, cardiovascular, metabolic, neurologic & psychiatric disorders, drug-resistant infections...

# Data citation model – top level



# Data Citation Generic Example

example of a data citation as it would appear *in a reference list*\*

**Principle 2: Credit and Attribution**  
(e.g. authors, repositories or other distributors and contributors)

**Principle 4: Unique Identifier** (e.g. DOI, Handle.). **Principle 5, 6 Access, Persistence:**  
A persistent link to a landing page with metadata and access information

Author(s), Year, Dataset Title, Data Repository or Archive, [Accession], Global Persistent Identifier, version or subset

**Principle 7: Version and granularity**  
(e.g. a version number or a query to a subset) In addition, access to versions or subsets should be available from the landing page,

\*Note that the format is not intended to be defined with this example, as formats will vary across publishers and communities [**Principle 8: Interoperability and flexibility**].

## ¬ (Global Persistent Identifier ⇒ DOI)

- Identifiers.org catalogs > 500 biomedical dbs
  - Very few of these use DOIs
  - Cannot require all 500 to assign DOIs as prerequisite for data citation
- Current standard practice:
  - Assign identifiers locally per db
  - Prefix-based global uniqueness e.g. “ENA:”

# Multiple provider problem

- Some dbs have multiple providers
  - Providers may add distinctive features
  - These providers need credit for their work
- Default: resolve to provider with best uptime
- Option: resolve to specific provider

# Multiple resolver problem

- Identifiers.org project at EBI
  - Resolve compact PIDs <prefix:accession>
  - Impressive other features
- N2T.net project at California Digital Library
  - Prefix-agnostic resolver: ARKs, DOIs, URNs, PMIDs, etc.
  - Suffix passthrough plus per-id and per-prefix resolver rules (hybrid approach)
- Can we harmonize these for global reliability?

 **Identifiers.org**

Examples: [ontology](#), [enzyme](#), [Japan](#), [EMBL](#)

Search

[Categories & tags](#)

[Home](#) | [Documentation](#) | [Services](#) | [About](#)

*Identifiers.org* is an established resolving system that enables the referencing of data for the scientific community, with a current focus on the Life Sciences domain. It handles persistent identifiers in the form of [URIs](#) and [CURIEs](#). This allows the referencing of data in both a location-independent and resource-dependent manner. The provision of resolvable identifiers ([URLs](#)) fits well with the [Semantic Web](#) vision, and the [Linked Data](#) initiative.

Connect with us

 [Twitter](#)

 [GitHub](#)

[Tweets by IdentifiersOrg](#)

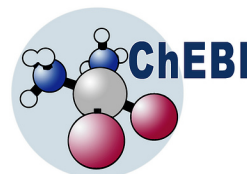
## Key facts

- **resolvable identifiers:** the system provides resolvable identifiers in the form of URIs and CURIEs
- **location independent:** the system is able to decouple the identification of records from the physical locations on the web where they can be retrieved
- **free to use:** there is no cost involved to register or use the identifiers or this facility
- **granularity of identifiers:** identifiers available at multiple levels (data collections, resources and data entities)
- **customisable behaviours:** the URI system can be tailored to needs in terms of formats available and preferred resolving locations
- **RDF support:** responses encoded in [RDF/XML](#) are provided via content negotiation, other formats are being added
- **community driven:** we rely upon the community to drive and direct development, as well as to contribute new data collections
- **curated resource:** dedicated curators and community feedback ensure that information in the Registry remains up-to-date and accurate
- **reliable:** there is an automated link monitoring system that checks registered resources daily
- **unrestricted scope:** currently focused on Life Sciences, but the scope is potentially unlimited (community feedback dependent)

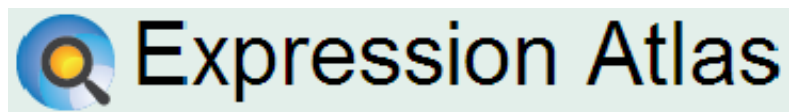
## Resolving mechanisms

A central, curated [registry](#) stores the information necessary to resolve either URI or CURIE forms of identifier, which are sent as queries to the Identifiers.org resolving system.

# Identifiers.org Registry





**550+ curated  
data collections**



# http://identifiers.org/registry

 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
 <a href="#">InChIKey</a>	inchikey	The IUPAC International Chemical Identifier (InChI, see MIR:00000383) is an identifier for chemical substances, and is derived solely from a structural representation of that substance. Since these can be quite unwieldy, particularly for web use, the InChIKey was developed. These are of a fixed length (25 character) and were created as a condensed, more web friendly, digital representation of the InChI.
<a href="#">Systems Biology Ontology</a>	sbo	The goal of the Systems Biology Ontology is to develop controlled vocabularies and ontologies tailored specifically for the kinds of problems being faced in Systems Biology, especially in the context of computational modeling. SBO is a project of the BioModels.net effort.
 <a href="#">Kyoto Encyclopedia of Genes and Genomes</a>	kegg	Kyoto Encyclopedia of Genes and Genomes (KEGG) is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies.
<a href="#">OMIM</a>	omim	Online Mendelian Inheritance in Man is a catalog of human genes and genetic disorders.
<a href="#">DOI</a>	doi	The Digital Object Identifier System is for identifying content objects in the digital environment.
<a href="#">RRID</a>	rrid	The Research Resource Identification Initiative provides RRDs to 4 main classes of resources: Antibodies, Cell Lines, Model Organisms, and Databases / Software tools.: Antibodies, Model Organisms, and Databases / Software tools. The initiative works with participating journals to intercept manuscripts in the publication process that use these resources, and allows publication authors to incorporate RRDs within the methods sections. It also provides resolver services that access curated data from 10 data sources: the antibody registry (a curated catalog of antibodies), the SciCrunch registry (a curated catalog of software tools and databases), and model organism nomenclature authority databases (MGI, FlyBase, WormBase, RGD), as well as various stock centers. These RRDs are aggregated and can be searched through SciCrunch.
<a href="#">Mammalian Phenotype Ontology</a>	mp	The Mammalian Phenotype Ontology (MP) classifies and organises phenotypic information related to the mouse and other mammalian species. This ontology has been applied to mouse phenotype descriptions in various databases allowing comparisons of data from diverse mammalian sources. It can facilitate in the identification of appropriate experimental disease models, and aid in the discovery of candidate disease genes and molecular signaling pathways.
<a href="#">Protein Modification Ontology</a>	mod	The Proteomics Standards Initiative modification ontology (PSI-MOD) aims to define a consensus nomenclature and ontology reconciling, in a hierarchical representation, the complementary descriptions of residue modifications.

# Identifiers.org URIs

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



[Data collection][Entity identifier]



<http://identifiers.org/taxonomy/9606>

## General information

Recommended name	<b>Taxonomy</b>
Alternative name(s)	NEWT NCBI taxonomy
Description	The taxonomy contains the relationships between all living forms for which nucleic acid or protein sequence have been determined.
Identifier pattern	^\\d+\$
Registry identifier	MIR:00000006

[taxonomy](#)

## Identification schemes

Namespace	taxonomy
URI	<a href="http://identifiers.org/taxonomy/">http://identifiers.org/taxonomy/</a>

### Alternative URI schemes

- <http://www.uniprot.org/taxonomy/>
- <http://purl.obolibrary.org/obo/NCBITaxon>
- <http://bio2rdf.org/taxonomy>
- urn:miriam:taxonomy

### Deprecated URI scheme(s)

## Physical locations (resources)

<b>Primary</b>	<a href="#">Resource</a> <a href="#">MIR:00100019</a>	Description	Taxonomy through UniProt PURL
		Access URL	<a href="http://purl.uniprot.org/taxonomy/\$id">http://purl.uniprot.org/taxonomy/\$id</a> [Example: <a href="#">9606</a> ]
		Institution	UniProt Consortium, USA, UK and Switzerland
		Website	<a href="http://www.uniprot.org/taxonomy/">http://www.uniprot.org/taxonomy/</a>
	<a href="#">Resource</a> <a href="#">MIR:00100299</a>	Description	European Nucleotide Archive (ENA)
		Access URL	<a href="http://www.ebi.ac.uk/ena/data/view/Taxon:\$id">http://www.ebi.ac.uk/ena/data/view/Taxon:\$id</a> [Example: <a href="#">9606</a> ]
		Institution	European Bioinformatics Institute, Hinxton, Cambridge, UK
		Website	<a href="http://www.ebi.ac.uk/ena/">http://www.ebi.ac.uk/ena/</a>
	<a href="#">Resource</a> <a href="#">MIR:00100007</a>	Description	NCBI Taxonomy
		Access URL	<a href="http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&amp;id=\$id">http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&amp;id=\$id</a> [Example: <a href="#">9606</a> ]
		Institution	National Center for Biotechnology Information, USA
		Website	<a href="http://www.ncbi.nlm.nih.gov/Taxonomy/">http://www.ncbi.nlm.nih.gov/Taxonomy/</a>
	<a href="#">Resource</a> <a href="#">MIR:00100695</a>	Description	Bio2RDF
		Access URL	<a href="http://taxonomy.bio2rdf.org/describe/?url=http://bio2rdf.org/taxonomy:\$id">http://taxonomy.bio2rdf.org/describe/?url=http://bio2rdf.org/taxonomy:\$id</a> [Example: <a href="#">9606</a> ]
		Institution	Bio2RDF.org
		Website	<a href="http://taxonomy.bio2rdf.org/ict/">http://taxonomy.bio2rdf.org/ict/</a>
		Description	BioPortal

## General information

Recommended name	<b>Taxonomy</b>
Alternative name(s)	NEWT NCBI taxonomy
Description	The taxonomy contains the relationships between all living forms for which nucleic acid or protein sequence have been determined.
Identifier pattern	^id+\$
Registry identifier	MIR:00000006

[taxonomy](#)

## Identification schemes

Namespace	taxonomy
URI	<a href="http://identifiers.org/taxonomy/">http://identifiers.org/taxonomy/</a>

### Alternative URI schemes

- <http://www.uniprot.org/taxonomy/>
- <http://purl.obolibrary.org/obo/NCBITaxon>
- <http://bio2rdf.org/taxonomy>
- urn:miriam:taxonomy

### Deprecated URI scheme(s)

## Physical locations (resources)

Primary	<a href="#">Resource</a> <a href="#">MIR:00100019</a>	Description	Taxonomy through UniProt Consortium
		Access URL	<a href="http://purl.uniprot.org/taxonomy/">http://purl.uniprot.org/taxonomy/</a>
		Institution	UniProt Consortium
		Website	<a href="http://www.uniprot.org/taxonomy/">http://www.uniprot.org/taxonomy/</a>
	<a href="#">Resource</a> <a href="#">MIR:00100299</a>	Description	European Nucleotide Archive
		Access URL	<a href="http://www.ebi.ac.uk/ena/">http://www.ebi.ac.uk/ena/</a>
		Institution	European Bioinformatics Institute, Hinxton, Cambridge, UK
		Website	<a href="http://www.ebi.ac.uk/ena/">http://www.ebi.ac.uk/ena/</a>
	<a href="#">Resource</a> <a href="#">MIR:00100007</a>	Description	NCBI Taxonomy
		Access URL	<a href="http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&amp;id=\$id">http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&amp;id=\$id</a> [Example: <a href="#">9606</a> ]
		Institution	National Center for Biotechnology Information, USA
		Website	<a href="http://www.ncbi.nlm.nih.gov/Taxonomy/">http://www.ncbi.nlm.nih.gov/Taxonomy/</a>
	<a href="#">Resource</a> <a href="#">MIR:00100695</a>	Description	Bio2RDF
		Access URL	<a href="http://taxonomy.bio2rdf.org/describe/?url=http://bio2rdf.org/taxonomy:\$id">http://taxonomy.bio2rdf.org/describe/?url=http://bio2rdf.org/taxonomy:\$id</a> [Example: <a href="#">9606</a> ]
		Institution	Bio2RDF.org
		Website	<a href="http://taxonomy.bio2rdf.org/fct/">http://taxonomy.bio2rdf.org/fct/</a>
		Description	BioPortal

## Identification schemes

Namespace	taxonomy
URI	<a href="http://identifiers.org/taxonomy/">http://identifiers.org/taxonomy/</a>

### Alternative URI schemes

- <http://www.uniprot.org/taxonomy/>
- <http://purl.obolibrary.org/obo/NCBITaxon>
- <http://bio2rdf.org/taxonomy>
- urn:miriam:taxonomy

### Deprecated URI scheme(s)

- <http://www.ncbi.nlm.nih.gov/Taxonomy/>
- <http://www.taxonomy.org/>

## General information

Recommended name	<b>Taxonomy</b>
Alternative name(s)	NEWT NCBI taxonomy
Description	The taxonomy contains the relationships between all living forms for which nucleic acid or protein sequence have been determined.
Identifier pattern	^\\d+\$
Registry identifier	MIR:00000006

[taxonomy](#)

## Identification schemes

Namespace	taxonomy
URI	<a href="http://identifiers.org/taxonomy/">http://identifiers.org/taxonomy/</a>

### Alternative URI schemes

- <http://www.uniprot.org/taxonomy/>
- <http://purl.obolibrary.org/obo/NCBITaxon>
- <http://bio2rdf.org/taxonomy>
- <urn:miriam:taxonomy>

### Deprecated URI scheme(s)

## Physical locations (resources)

<b>Primary</b>	<a href="#">Resource</a> <a href="#">MIR:00100019</a>	Description	Taxonomy through UniProt PURL
		Access URL	<a href="http://purl.uniprot.org/taxonomy/\$id">http://purl.uniprot.org/taxonomy/\$id</a> [Example: <a href="#">9606</a> ]
		Institution	UniProt Consortium, USA, UK and Switzerland
		Website	<a href="http://www.uniprot.org/taxonomy/">http://www.uniprot.org/taxonomy/</a>
	<a href="#">Resource</a> <a href="#">MIR:00100299</a>	Description	European Nucleotide Archive (ENA)
		Access URL	<a href="http://www.ebi.ac.uk/ena/data/view/Taxon:\$id">http://www.ebi.ac.uk/ena/data/view/Taxon:\$id</a> [Example: <a href="#">9606</a> ]
		Institution	European Bioinformatics Institute, Hinxton, Cambridge, UK
		Website	<a href="http://www.ebi.ac.uk/ena/">http://www.ebi.ac.uk/ena/</a>
	<a href="#">Resource</a> <a href="#">MIR:00100007</a>	Description	NCBI Taxonomy
		Access URL	<a href="http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&amp;id=\$id">http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&amp;id=\$id</a> [Example: <a href="#">9606</a> ]
		Institution	National Center for Biotechnology Information, USA
		Website	<a href="http://www.ncbi.nlm.nih.gov/Taxonomy/">http://www.ncbi.nlm.nih.gov/Taxonomy/</a>
	<a href="#">Resource</a> <a href="#">MIR:00100695</a>	Description	Bio2RDF
		Access URL	<a href="http://taxonomy.bio2rdf.org/describe/?url=http://bio2rdf.org/taxonomy:\$id">http://taxonomy.bio2rdf.org/describe/?url=http://bio2rdf.org/taxonomy:\$id</a> [Example: <a href="#">9606</a> ]
		Institution	Bio2RDF.org
		Website	<a href="http://taxonomy.bio2rdf.org/fct/">http://taxonomy.bio2rdf.org/fct/</a>
		Description	BioPortal

# Health check history

Health history of: MIR:00100651 (dbEST, dbEST through DNA Data Bank of Japan (DDBJ)).

2016

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
January																															
February																															
March																															
April																															
May																															
June																															
July																															
August																															
September																															
October																															
November																															
December																															

2015

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
January																															
February																															
March																															
April																															
May																															
June																															
July																															
August																															
September																															
October																															
November																															
December																															

# Development considerations

- Recognize user preferences
  - allow **users** to express resolution preferences
  - allow **format** preferences
- Acknowledge data providers
  - record **primary** resource
  - **used by default**
- Community curation & accuracy
  - Allow **data providers** to maintain their own records
  - Explore automated procedures to harvest
- More resources mint their own perennial URIs
  - record URIs, provide **conversion** facility
- Provide/harvest metadata for resources and their records

# The N2T.net resolver

---

Born in 2007, the N2T (Name-to-Thing) resolver was specifically made to avoid another identifier silo

- Scheme-agnostic: services any kind of name
- Initially provided resolution for ARK, DOI, URN, PMID, and a handful of other schemes
- Primary in California, first replica in Edinburgh
- Conceived for community/consortial ownership
  - Creator and first administrator: California Digital Library

# Libraries are domain-agnostic

---

Why build a global N2t resolver at the California Digital Library?

- Huge institution with huge range of disciplines
- Dependent on many kinds of identifiers
- Already maintaining the ARK identifier scheme



# Need for generic identifier services

---

- EZID system for creating and managing ids (logged UI, API)
- Separate N2T system for public access *resolution*



N2T (Name-to-Thing) resolves with hybrid approach

- Stored rules: to redirect to other ARK and other scheme resolvers
- Stored ids: 6M EZID ids and 12M Internet Archive ids
  - Demonstrated resolution with 61M Crossref DOIs

# http://identifiers.org/registry

 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
 <a href="#">InChIKey</a>	inchikey	The IUPAC International Chemical Identifier (InChI, see MIR:00000383) is an identifier for chemical substances, and is derived solely from a structural representation of that substance. Since these can be quite unwieldy, particularly for web use, the InChIKey was developed. These are of a fixed length (25 character) and were created as a condensed, more web friendly, digital representation of the InChI.
<a href="#">Systems Biology Ontology</a>	sbo	The goal of the Systems Biology Ontology is to develop controlled vocabularies and ontologies tailored specifically for the kinds of problems being faced in Systems Biology, especially in the context of computational modeling. SBO is a project of the BioModels.net effort.
 <a href="#">Kyoto Encyclopedia of Genes and Genomes</a>	kegg	Kyoto Encyclopedia of Genes and Genomes (KEGG) is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies.
<a href="#">OMIM</a>	omim	Online Mendelian Inheritance in Man is a catalog of human genes and genetic disorders.
<a href="#">DOI</a>	doi	The Digital Object Identifier System is for identifying content objects in the digital environment.
<a href="#">RRID</a>	rrid	The Research Resource Identification Initiative provides RRDs to 4 main classes of resources: Antibodies, Cell Lines, Model Organisms, and Databases / Software tools.: Antibodies, Model Organisms, and Databases / Software tools. The initiative works with participating journals to intercept manuscripts in the publication process that use these resources, and allows publication authors to incorporate RRDs within the methods sections. It also provides resolver services that access curated data from 10 data sources: the antibody registry (a curated catalog of antibodies), the SciCrunch registry (a curated catalog of software tools and databases), and model organism nomenclature authority databases (MGI, FlyBase, WormBase, RGD), as well as various stock centers. These RRDs are aggregated and can be searched through SciCrunch.
<a href="#">Mammalian Phenotype Ontology</a>	mp	The Mammalian Phenotype Ontology (MP) classifies and organises phenotypic information related to the mouse and other mammalian species. This ontology has been applied to mouse phenotype descriptions in various databases allowing comparisons of data from diverse mammalian sources. It can facilitate in the identification of appropriate experimental disease models, and aid in the discovery of candidate disease genes and molecular signaling pathways.
<a href="#">Protein Modification Ontology</a>	mod	The Proteomics Standards Initiative modification ontology (PSI-MOD) aims to define a consensus nomenclature and ontology reconciling, in a hierarchical representation, the complementary descriptions of residue modifications.

# Namespace prefix

## 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
 <a href="#">InChIKey</a>	inchikey	The IUPAC International Chemical Identifier (InChI, see MIR:00000383) is an identifier for chemical substances, and is derived solely from a structural representation of that substance. Since these can be quite unwieldy, particularly for web use, the InChIKey was developed. These are of a fixed length (25 character) and were created as a condensed, more web friendly, digital representation of the InChI.
<a href="#">Systems Biology Ontology</a>	sbo	The goal of the Systems Biology Ontology is to develop controlled vocabularies and ontologies tailored specifically for the kinds of problems being faced in systems biology.
 <a href="#">Kyoto Encyclopedia of Genes and Genomes</a>	kegg	Kyoto Encyclopedia of Genes and Genomes (KEGG) is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genomics technologies.
<a href="#">OMIM</a>	omim	Online Mendelian Inheritance in Man (OMIM) is a catalog of human genes and genetic disorders.
<a href="#">DOI</a>	doi	The Digital Object Identifier System is for identifying content objects in the digital environment.
<a href="#">RRID</a>	rrid	The Research Resource Identification Initiative provides RRIDs to 4 main classes of resources: Antibodies, Cell Lines, Model Organisms, and Databases / Software tools. The initiative works with participating journals to intercept manuscripts in the publication process that use these resources, and allows publication authors to incorporate RRIDs within the methods section of their manuscripts. Services that access curated data from 10 data sources: the antibody registry (a curated catalog of antibodies), the SIB database (a curated catalog of software tools and databases), and model organism nomenclature authority databases (MGI, FlyBase, RGD, etc.) stock centers. These RRIDs are aggregated and can be searched through SciCrunch.
<a href="#">Mammalian Phenotype Ontology</a>	mp	The Mammalian Phenotype Ontology (MP) is a controlled vocabulary for phenotypic information related to the mouse and other mammalian species. This ontology has been developed by integrating information from various databases allowing comparisons of data from diverse mammalian sources. It can facilitate in the identification of appropriate experimental disease models, and aid in the discovery of candidate disease genes and molecular signaling pathways.
<a href="#">Protein Modification Ontology</a>	mod	The Proteomics Standards Initiative modification ontology (PSI-MOD) aims to define a consensus nomenclature and ontology reconciling, in a hierarchical representation, the complementary descriptions of residue modifications.

- **Namespace prefix implementation**

- Namespace used as prefix

- GO:0006915

- PDB:2gc4

- **YAML file**

- N2t ↔ [identifiers.org](http://identifiers.org)

- <http://identifiers.org/GO:0006915>

- <http://n2t.net/GO:0006915>

## General information

Recommended name	<b>Taxonomy</b>
Alternative name(s)	NEWT NCBI taxonomy
Description	The taxonomy contains the relationships between all living forms for which nucleic acid or protein sequence have been determined.
Identifier pattern	^\\d+\$
Registry identifier	MIR:00000006

[taxonomy](#)

## Identification schemes

Namespace	taxonomy
URI	<a href="http://identifiers.org/taxonomy/">http://identifiers.org/taxonomy/</a>

### Alternative URI schemes

- <http://www.uniprot.org/taxonomy/>
- <http://purl.obolibrary.org/obo/NCBITaxon>
- <http://bio2rdf.org/taxonomy>
- urn:miriam:taxonomy

### Deprecated URI scheme(s)

## Physical locations (resources)

<b>Primary</b>	<a href="#">Resource</a> <a href="#">MIR:00100019</a>	Description	Taxonomy through UniProt PURL
		Access URL	<a href="http://purl.uniprot.org/taxonomy/\$id">http://purl.uniprot.org/taxonomy/\$id</a> [Example: <a href="#">9606</a> ]
		Institution	UniProt Consortium, USA, UK and Switzerland
		Website	<a href="http://www.uniprot.org/taxonomy/">http://www.uniprot.org/taxonomy/</a>
	<a href="#">Resource</a> <a href="#">MIR:00100299</a>	Description	European Nucleotide Archive (ENA)
		Access URL	<a href="http://www.ebi.ac.uk/ena/data/view/Taxon:\$id">http://www.ebi.ac.uk/ena/data/view/Taxon:\$id</a> [Example: <a href="#">9606</a> ]
		Institution	European Bioinformatics Institute, Hinxton, Cambridge, UK
		Website	<a href="http://www.ebi.ac.uk/ena/">http://www.ebi.ac.uk/ena/</a>
	<a href="#">Resource</a> <a href="#">MIR:00100007</a>	Description	NCBI Taxonomy
		Access URL	<a href="http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&amp;id=\$id">http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&amp;id=\$id</a> [Example: <a href="#">9606</a> ]
		Institution	National Center for Biotechnology Information, USA
		Website	<a href="http://www.ncbi.nlm.nih.gov/Taxonomy/">http://www.ncbi.nlm.nih.gov/Taxonomy/</a>
	<a href="#">Resource</a> <a href="#">MIR:00100695</a>	Description	Bio2RDF
		Access URL	<a href="http://taxonomy.bio2rdf.org/describe/?url=http://bio2rdf.org/taxonomy:\$id">http://taxonomy.bio2rdf.org/describe/?url=http://bio2rdf.org/taxonomy:\$id</a> [Example: <a href="#">9606</a> ]
		Institution	Bio2RDF.org
		Website	<a href="http://taxonomy.bio2rdf.org/ict/">http://taxonomy.bio2rdf.org/ict/</a>
		Description	BioPortal

### General information

Recommended name	Taxonomy
Alternative name(s)	NEWT
URI	http://purl.obolibrary.org/obo/ncbi/taxonomy/
Alternative URI(s)	<ul style="list-style-type: none"> <li>http://www.uniprot.org/taxonomy/</li> <li>http://purl.ebi.ac.uk/obo/NCBITaxon</li> <li>http://bio2rdf.org/taxonomy</li> <li>um.miriam.taxonomy</li> </ul>
Deprecated URI scheme(s)	

### Provider code implementation

- Assign a code to providers
  - NCBI, EBI, ..
  - OLS, BPTL, ..
  - PDBj, PDBe, RCSB, ...

- Identifiers.org
  - ncbi/taxonomy:9606
  - amigo/go:0006915
- n2t
  - amigo/go:0006915

### Physical locations (resources)

Primary	Resource MIR:00100019	Description	Taxonomy through UniProt PURL
		Access URL	<a href="http://purl.uniprot.org/taxonomy/\$id">http://purl.uniprot.org/taxonomy/\$id</a> [Example: 9606]
		Institution	UniProt Consortium, USA, UK and Switzerland
		Website	<a href="http://www.uniprot.org/taxonomy/">http://www.uniprot.org/taxonomy/</a>
	Resource MIR:00100299	Description	European Nucleotide Archive (ENA)
		Access URL	<a href="http://www.ebi.ac.uk/ena/data/view/Taxon:\$id">http://www.ebi.ac.uk/ena/data/view/Taxon:\$id</a> [Example: 9606]
		Institution	European Bioinformatics Institute, Hinxton, Cambridge, UK
		Website	<a href="http://www.ebi.ac.uk/ena/">http://www.ebi.ac.uk/ena/</a>
	Resource MIR:00100007	Description	NCBI Taxonomy
		Access URL	<a href="http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&amp;id=\$id">http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&amp;id=\$id</a> [Example: 9606]
		Institution	National Center for Biotechnology Information, USA
		Website	<a href="http://www.ncbi.nlm.nih.gov/Taxonomy/">http://www.ncbi.nlm.nih.gov/Taxonomy/</a>
	Resource MIR:00100695	Description	Bio2RDF
		Access URL	<a href="http://taxonomy.bio2rdf.org/describe?url=http://bio2rdf.org/taxonomy:\$id">http://taxonomy.bio2rdf.org/describe?url=http://bio2rdf.org/taxonomy:\$id</a> [Example: 9606]
		Institution	Bio2RDF.org
		Website	<a href="http://taxonomy.bio2rdf.org/ict/">http://taxonomy.bio2rdf.org/ict/</a>
		Description	BioPortal

# YAML

## Alignment of namespace prefixes

```
223 namespace_prefix: pdb
224 provider_prefix: rcsb
225 description: RCSB PDB
226 homepage: http://www.pdb.org/
227 -
228 namespace_prefix: pdb
229 provider_prefix: pdbe
230 description: Protein Databank in Europe (PDBe)
231 homepage: http://www.pdbe.org/
232 -
233 namespace_prefix: pdb
234 provider_prefix:
235 description: Proteopedia
236 homepage: http://www.proteopedia.org/
237 -
238 namespace_prefix: pdb
239 provider_prefix: pdbj
240 description: Protein Data Bank Japan (PDBj)
241 homepage: http://www.pdbj.org/
242 -
243 namespace_prefix: pdb
244 provider_prefix:
245 description: Protein Databank through PDBsum
246 homepage: http://www.ebi.ac.uk/pdbsum/
247 -
248 namespace_prefix: go
249 provider_prefix: ebi
250 description: QuickGO (Gene Ontology browser)
251 homepage: http://www.ebi.ac.uk/QuickGO/
252 -
253 namespace_prefix: go
254 provider_prefix: amigo
255 description: AmiGO 2
256 homepage: http://amigo.geneontology.org/
```

<https://github.com/identifiers-org/prefix/blob/master/prefix.yaml>

# YAML

Alignment of  
namespace prefixes

Alignment of  
provider codes

```
223 namespace_prefix: pdb
224 provider_prefix: rcsb
225 description: RCSB PDB
226 homepage: http://www.pdb.org/
227 -
228 namespace_prefix: pdb
229 provider_prefix: pdbe
230 description: Protein Databank in Europe (PDBe)
231 homepage: http://www.pdbe.org/
232 -
233 namespace_prefix: pdb
234 provider_prefix:
235 description: Proteopedia
236 homepage: http://www.proteopedia.org/
237 -
238 namespace_prefix: pdb
239 provider_prefix: pdbj
240 description: Protein Data Bank Japan (PDBj)
241 homepage: http://www.pdbj.org/
242 -
243 namespace_prefix: pdb
244 provider_prefix:
245 description: Protein Databank through PDBsum
246 homepage: http://www.ebi.ac.uk/pdbsum/
247 -
248 namespace_prefix: go
249 provider_prefix: ebi
250 description: QuickGO (Gene Ontology browser)
251 homepage: http://www.ebi.ac.uk/QuickGO/
252 -
253 namespace_prefix: go
254 provider_prefix: amigo
255 description: AmiGO 2
256 homepage: http://amigo.geneontology.org/
```

<https://github.com/identifiers-org/prefix/blob/master/prefix.yaml>

# Prefix alignment

- Initial implementation
  - representative
  - popular databases
- Helped derive some simple rules
  - Alphanumeric prefixes, (period and underscore allowed)
- Intricacies:
- (Identifiers.org) namespace prefix condensation rules
  - eg. ontologies, GO:GO:0006915 vs. GO:0006915
- Legacy prefix consideration (n2t ↔ identifiers.org)
  - ‘Aliasing’ eg. ‘pmid’ ↔ ‘pubmed’, ‘taxon’ ↔ ‘taxonomy’
  - n2t supports both forms

# Acknowledgements

## EBI Team

Nick Juty  
Sarala Wimalaratne  
Henning Hermjakob  
Nicolas Le Novère

## CDL Team

John Kunze  
Greg Janee  
Joan Starr

## DCIP/Force11

Tim Clark (Harvard)  
Jeff Grethe (UCSD)  
Raphael Jiminez (ELIXIR)  
Niall Beard (Manchester)  
Julie McMurray (UCSB)  
Ian Fore (NCI/NIH)

## Contact

[sarala@ebi.ac.uk](mailto:sarala@ebi.ac.uk)  
[jak@ucop.edu](mailto:jak@ucop.edu)

