Assign locally, resolve globally: prefix-based collection access

Tim Clark, Nick Juty, John Kunze

PIDapalooza
9th-10th 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

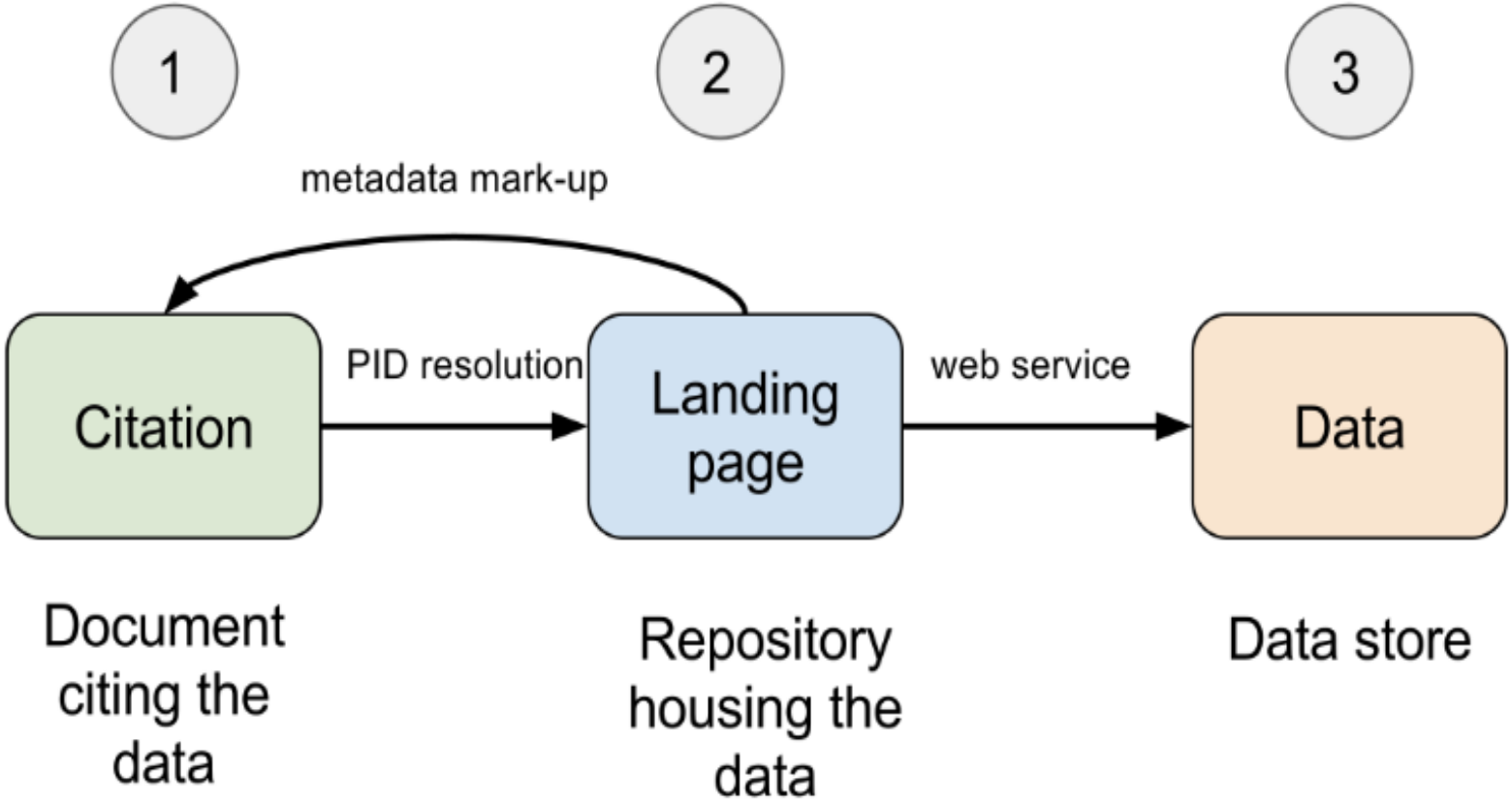
The Data Citation Use Case

- 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?

http://identifiers.org

Identifiers.org

Search

Examples: ontology, enzyme, Japan, EMBL

Categories & tags

Home | Documentation | Services | About

Identifiers.org is an established resolving system the 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.

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.

Connect with us

Twitter GitHub

Tweets by IdentifiersOrg

Identifiers.org Registry

DrugBank
Open Data Drug & Drug Target Database

ChEBI

Reactome

GO
the Gene Ontology

550+ curated data collections

UniProt

PubMed

ChEMBL

Expression Atlas

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
InChIKey	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 unwieldly, 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.
Systems Biology Ontology	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.
Kyoto Encyclopedia of Genes and Genomes	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.
OMIM	omim	Online Mendelian Inheritance in Man is a catalog of human genes and genetic disorders.
DOI	doi	The Digital Object Identifier System is for identifying content objects in the digital environment.
RRID	rrid	The Research Resource Identification Initiative provides RRIDs 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 RRIDs 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 RRIDs are aggregated and can be searched through SciCrunch.
Mammalian Phenotype Ontology	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.
Protein Modification Ontology	mod	The Proteomics Standards Initiative modification ontology (PSI-MOD) aims to define a concensus 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

Data collection: *Taxonomy*

Overview | Miscellaneous | RDF/XML | Turtle

General information

taxonomy

Recommended name	**Taxonomy**
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

Identification schemes

Namespace	taxonomy
URI	http://identifiers.org/taxonomy/

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)

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

Data collection: *Taxonomy*

Overview | Miscellaneous | RDF/XML | Turtle

General information

taxonomy

Recommended name	**Taxonomy**
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

Identification schemes

Namespace	taxonomy
URI	http://identifiers.org/taxonomy/

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/

Physical locations (resources)

	Resource MIR:00100019	Description	Taxonomy thro
		Access URL	http://purl.uni
		Institution	UniProt Consor
		Website	http://www.un
	Resource MIR:00100299	Description	European Nu
		Access URL	http://www.ebi
		Institution	European Bioinformatics Institute, Hinxton, Cambridge, UK
		Website	http://www.ebi.ac.uk/ena/
Primary	Resource MIR:00100007	Description	NCBI Taxonomy
		Access URL	http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=$id [Example: 9606]
		Institution	National Center for Biotechnology Information, USA
		Website	http://www.ncbi.nlm.nih.gov/Taxonomy/
	Resource MIR:00100695	Description	Bio2RDF
		Access URL	http://taxonomy.bio2rdf.org/describe/?url=http://bio2rdf.org/taxonomy:$id [Example: 9606]
		Institution	Bio2RDF.org
		Website	http://taxonomy.bio2rdf.org/fct/
		Description	BioPortal

Data collection: *Taxonomy*

Overview | Miscellaneous | RDF/XML | Turtle

General information

taxonomy

Recommended name	**Taxonomy**
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

Identification schemes

Namespace	taxonomy
URI	http://identifiers.org/taxonomy/

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)

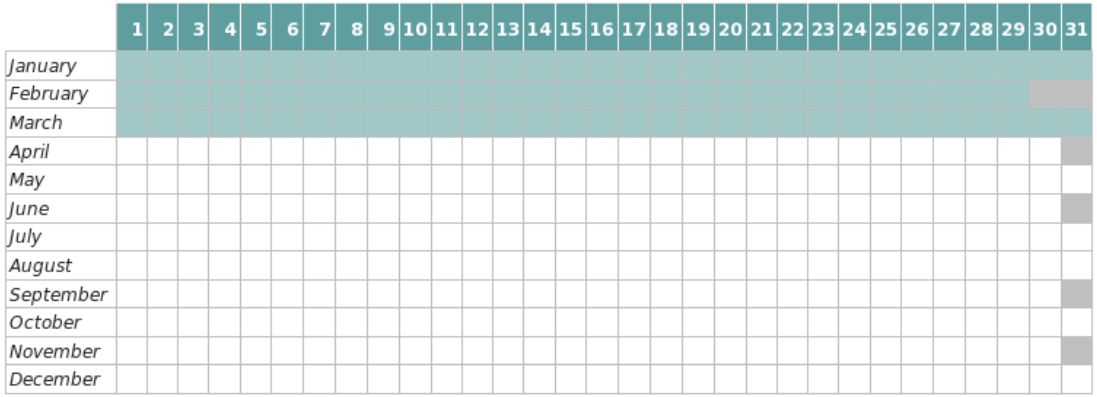
Physical locations (resources)

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

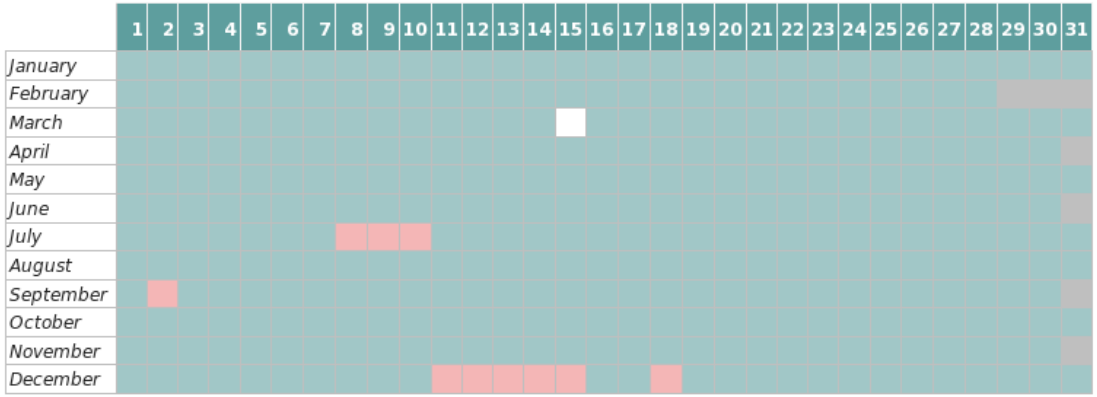
Health check history

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

2016



2015



EMBL-EBI

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/consoritial ownership
  - Creator and first administrator: California Digital Library

University of California
CDL
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

University of California
CDL
California Digital Library

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
InChIKey	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 unwieldly, 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.
Systems Biology Ontology	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.
Kyoto Encyclopedia of Genes and Genomes	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.
OMIM	omim	Online Mendelian Inheritance in Man is a catalog of human genes and genetic disorders.
DOI	doi	The Digital Object Identifier System is for identifying content objects in the digital environment.
RRID	rrid	The Research Resource Identification Initiative provides RRIDs 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 RRIDs 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 RRIDs are aggregated and can be searched through SciCrunch.
Mammalian Phenotype Ontology	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.
Protein Modification Ontology	mod	The Proteomics Standards Initiative modification ontology (PSI-MOD) aims to define a concensus 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
InChIKey	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 unwieldly, 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.
Systems Biology Ontology	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 [illegible] the BioModels.net effort.
Kyoto Encyclopedia of Genes and Genomes	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 [illegible] other high-throughput experimental technologies.
OMIM	omim	Online Mendelian Inheritance [illegible] catalog of human genes and genetic disorders.
DOI	doi	The Digital Object Identifier System is for identifying content objects in the digital environment.
RRID	rrid	The Research Resource Identification Initiative provides RRIDs 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 RRIDs within the methods section [illegible] services that access curated data from 10 data sources: the antibody registry (a curated catalog of antibodies), the [illegible] (a curated catalog of software tools and databases), and model organism nomenclature authority databases (MGI, [illegible] stock centers. These RRIDs are aggregated and can be searched through SciCrunch.
Mammalian Phenotype Ontology	mp	The Mammalian Phenotype [illegible] related to the mouse and other mammalian species. This ontology [illegible] ious 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.
Protein Modification Ontology	mod	The Proteomics Standards Initiative modification ontology (PSI-MOD) aims to define a concensus 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/GO:0006915
    - http://n2t.net/GO:0006915

Data collection: *Taxonomy*

Overview | Miscellaneous | RDF/XML | Turtle

General information

taxonomy

Recommended name	**Taxonomy**
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

Identification schemes

Namespace	taxonomy
URI	http://identifiers.org/taxonomy/

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)

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

Provider codes

- 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)

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

YAML

Alignment of namespace prefixes

```yaml
    namespace_prefix: pdb
    provider_prefix: rcsb
    description: RCSB PDB
    homepage: http://www.pdb.org/
-
    namespace_prefix: pdb
    provider_prefix: pdbe
    description: Protein Databank in Europe (PDBe)
    homepage: http://www.pdbe.org/
-
    namespace_prefix: pdb
    provider_prefix:
    description: Proteopedia
    homepage: http://www.proteopedia.org/
-
    namespace_prefix: pdb
    provider_prefix: pdbj
    description: Protein Data Bank Japan (PDBj)
    homepage: http://www.pdbj.org/
-
    namespace_prefix: pdb
    provider_prefix:
    description: Protein Databank through PDBsum
    homepage: http://www.ebi.ac.uk/pdbsum/
-
    namespace_prefix: go
    provider_prefix: ebi
    description: QuickGO (Gene Ontology browser)
    homepage: http://www.ebi.ac.uk/QuickGO/
-
    namespace_prefix: go
    provider_prefix: amigo
    description: AmiGO 2
    homepage: http://amigo.geneontology.org/
```

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

YAML

Alignment of namespace prefixes

Alignment of provider codes

```yaml
    namespace_prefix: pdb
    provider_prefix: rcsb
    description: RCSB PDB
    homepage: http://www.pdb.org/
-
    namespace_prefix: pdb
    provider_prefix: pdbe
    description: Protein Databank in Europe (PDBe)
    homepage: http://www.pdbe.org/
-
    namespace_prefix: pdb
    provider_prefix:
    description: Proteopedia
    homepage: http://www.proteopedia.org/
-
    namespace_prefix: pdb
    provider_prefix: pdbj
    description: Protein Data Bank Japan (PDBj)
    homepage: http://www.pdbj.org/
-
    namespace_prefix: pdb
    provider_prefix:
    description: Protein Databank through PDBsum
    homepage: http://www.ebi.ac.uk/pdbsum/
-
    namespace_prefix: go
    provider_prefix: ebi
    description: QuickGO (Gene Ontology browser)
    homepage: http://www.ebi.ac.uk/QuickGO/
-
    namespace_prefix: go
    provider_prefix: amigo
    description: AmiGO 2
    homepage: http://amigo.geneontology.org/
```

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

Prefix alignment

- Initial implemention
  - 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

DCIP Identifiers Workshop, June 2, 2016, Harvard University, Cambridge MA
John Kunze (CDL), Niall Beard (Manchester), Tim Clark (Harvard), Nick Juty (EBI), Ian Fore (NIH), Julie McMurry (UCSB), Jeff Grethe (UCSD), Rafael Jimenez (ELIXIR), Sarala Wimalaratne (EBI)

Contact

sarala@ebi.ac.uk
jak@ucop.edu