Chapter 12

Controlled Annotations for Systems Biology

Nick Juty, Camille Laibe, and Nicolas Le Novère

Abstract

The aim of this chapter is to provide sufficient information to enable a reader, new to the subject of Systems Biology, to create and use effectively controlled annotations, using resolvable Identifiers.org Uniform Resource Identifiers (URIs). The text details the underlying requirements that have led to the development of such an identification scheme and infrastructure, the principles that underpin its syntax and the benefits derived through its use. It also places into context the relationship with other standardization efforts, how it differs from other pre-existing identification schemes, recent improvements to the system, as well as those that are planned in the future. Throughout, the reader is provided with explicit examples of use and directed to supplementary information where necessary.

1 Introduction on MIRIAM Guidelines

Typically models generated in the latter part of the twentieth century were created in isolation, usually by small groups or by individuals. They were frequently encoded in custom formats or were directly written in a programming languages, contained non-standard terminologies to describe model components, and were often simulated or processed with proprietary software applications. Together, these factors resulted in a largely unusable body of work; since models could not be shared with other groups (custom formats), it was not clear what was being modeled (nonstandard nomenclature with insufficient metadata), or the simulation results could not be repeated (software specificity or unavailability).

Over the past decade, a number of standardization efforts have risen to address these deficiencies. There are now a number of description formats, largely based on XML (eXtensible Markup Language) *see* **Note 1**, which are suitable for the representation of models. These include, for instance, Systems Biology Markup Language (SBML) *see* **Note 2,** [1]. In addition, many other formats can be converted into a standardized representation, such as SBML, through the use of community-developed software.

Maria Victoria Schneider (ed.), *In Silico Systems Biology*, Methods in Molecular Biology, vol. 1021,
DOI 10.1007/978-1-62703-450-0_12,

To facilitate the harmonization of terminologies used in mathematical modeling, there now exists a cornucopia of ontologies, which themselves are developed according to shared community guidelines (Open Biomedical Ontologies Foundry (*see* **Note 3,** ref. [2])). These ontologies can be used, for example, to define the roles of various model components and the mathematical equations that describe their behaviors (see Systems Biology Ontology (*see* **Note 4,** ref. [3])), or to describe the algorithms that are needed to reproduce previously demonstrated simulation results (see KiSAO, Kinetic Simulation Algorithm Ontology (*see* **Note 5,** ref. [3])).

Various communities across the biological sciences also define their own Minimum Information checklists (MIs), specifying the key information that should be included with their (experimental) data to aid in their reuse (MIBBI, Minimum Information for Biological and Biomedical Investigations) (*see* **Note 6,** ref. [4])). In the field of Systems Biology, this yielded the Minimum Information Required in the Annotation of Models (MIRIAM; *see* ref. [5])).

The MIRIAM Guidelines are a community-developed effort to define a minimal set of information that should be provided within a model. This information should be sufficient to enable a model to be reused in the manner intended by its creator and is formalized as a set of guidelines to which a model must adhere to be deemed MIRIAM-compliant.

The MIRIAM Guidelines are composed of three sections, each dealing with a different aspect of a model and the manner in which it is encoded: *reference correspondence*, *attribution annotation*, and *external resource annotation*. Briefly, the *reference correspondence* section details information relating to the file format of the model, the accuracy with which it reflects the (biological) process under consideration, and its instantiability in a simulation. The *attribution annotation* section deals with information pertaining to the model creation process, its modification, and the terms under which it can be (re)distributed. The interested reader should consult the original publication for further details regarding these components of the Guidelines [5].

External resource annotation, the final section of the MIRIAM Guidelines, describes how to formalize the relationships between model components, and information about those components that is held externally, for instance on the World Wide Web. The objective of this final section of the Guidelines is to ensure that this information, or metadata, is constructed in such a manner as to prolong its longevity and accuracy. The following section details the considerations that were made in addressing this final part of the Guidelines, providing information on "metadata" and the concepts and existing frameworks that were leveraged to address the highlighted issues. The detailed requirements to comply with this section of the Guidelines, together with examples of use, follow in the section entitled "External resource annotation."

2 Metadata and Annotation

Metadata is often, and vaguely, defined as "data about data" and may refer to some information held elsewhere, perhaps in a repository or database, that relates to or sheds light on the present subject. Annotation is the process by which, in some shorthand notation, one can provide the "reader" with or direct him to this additional information. Annotations can be thought of as supplementary information which can be used to assist in clarification or definition of data components, but are not themselves required in the processing of that data. For example, in the context of modeling, the annotations provided within a model are not necessary to run a simulation.

Annotations can take many forms, many of which are not suitable for formal use. Referred to as "uncontrolled" annotations, they may be expressed as raw text, directly copy/pasted from the information source or web page address, or simply cite an identifier from a database, presented without context. These contribute to many downstream issues such as their unsuitability for computational processing, their unreliability due to fragility and changeability of web pages, and their ambiguities, respectively.

Identifiers assigned to data sets by their providers are almost exclusively composed from a limited pool of characters (alphanumeric). It is therefore often the case that an identifier from one data set is also a legitimate and valid identifier for a completely unrelated piece of information from a different data provider. For instance, the identifier "9606" describes *Homo sapiens* in the NCBI Taxonomy (*see* **Note 7**), a species of bird (*Bombycilla cedrorum*) in the BOLD taxonomy (*see* **Note 8**), and a German article in PubMed (*see* **Note 9**).

Even when care is taken to identify data using stable and established sources, there are some rare instances, in which an identifier scheme can be superseded. Table 1 illustrates the changes implemented in, what was at the time known as "EMBL bank," but is now known as the European Nucleotide Archive (ENA) (*see* **Note 10**).

Table 1
The history of protein identification syntax by release of the European Nucleotide archive

EMBL bank release (month/year)	Protein identification
43 (06/1995)	/note="pid:g2285"
45 (12/1995)	/db_xref="PID:g2285"
58 (03/1999)	/protein_id="CAA03857.1"

Table 2
The web addresses listed all provide alternative means to access exactly the same information from the Enzyme Nomenclature (http://www.chem.qmul.ac.uk/iubmb/enzyme/)

http://www.enzyme-database.org/query.php?ec=1.1.1.1
http://www.genome.jp/dbget-bin/www_bget?ec:1.1.1.1
htttp://www.ebi.ac.uk/intenz/query?cmd=SearchEC&ec=1.1.1.1
http://enzyme.expasy.org/EC/1.1.1.1

In some cases, the change in the syntax used by data providers can be subtle, as highlighted by the change in letter-case of "pid" between release 43 and 45 (Table 1), in ENA. Though not frequent, an entire identifier scheme can itself be deprecated in favor of, or subsumed into, an alternative scheme. Such a transition is shown in release 58 above (PID). In such circumstances, data providers should provide a mapping service to such entries, lest they be lost.

Databases are often accessed through a query-able interface. The resultant web address displayed is usually linkable in that it can be copied and pasted as text. However, web addresses often specify intrinsically an adopted architecture, specify a retrieval system, or direct one to a specific resolving location. Hence, with data being mirrored in various geographical locations, the copying of simple web addresses restricts one to a specific resource. If the specified resource is "down" at the time of query, relevant information cannot be accessed. In addition, over time, some URLs may also become obsolete. Some examples of different web addresses that provide exactly the same information are shown in Table 2.

An additional complication arises from database nomenclature itself. Identifiers provided by the Universal Protein Resource (UniProt (*see* **Note 11**)) have previously been known as "SWISS-PROT," "UniProt/Swiss-Prot," "UniProtKB/Swiss-Prot," and "UNP" identifiers. While self-evident to the reader, particularly one grounded in the biological sciences, the computational processing of such names, together with the diversity of associated web addresses, can be problematic and error-prone.

Clearly the use of "raw" text or any one of a plethora of web addresses, as the basis of an annotation, makes them short-lived, fragile, and difficult to process. Consequently, since the incorporation of annotations within a model has numerous benefits, a "controlled" metadata provision methodology is required.

2.1 Controlled Annotations

Controlled annotations are those which follow a defined structure and syntax. These need to address the major issues highlighted above, namely a way to handle the nomenclature used to identify

a set of data (SWISS-PROT vs. UniProtKB/Swiss-Prot), and a way to refer to a piece of data regardless of the architecture through which it is resolved, or of its geographical location (databases query mechanisms and remotely mirrored data). In the field of computer science, such mechanisms already exist: *Uniform Resource Identifiers* (URIs).

2.2 Uniform Resource Identifiers and Namespaces

A Uniform Resource Identifier (URI) (*see* **Note 12**) is a string of characters that is used to identify a resource and comes in two forms: Uniform Resource Name (URN) (*see* **Note 13**) and Uniform Resource Location (URL). Most people will be familiar with URLs; however, there is a key difference between the two which lies in the fact that a URN specifies only a name for a resource, while a URL specifies a name as well as a resolving location.

A namespace is a set of reserved strings of characters that are used to uniquely and unambiguously identify a pool of information. For example, the set of data available from the "Transport Classification Database" (*see* **Note 14**) is assigned the namespace *tcdb*.

By combining the use of a namespace with identifiers supplied by data providers in a URI, it is possible to build unique, robust, and perennial identifiers. To enable such identifiers to be used within any given community, and to ensure that they are used consistently, it is necessary to design a common syntax for encoding identifiers (URIs) and to share a list of legitimate namespaces. In our case, this list of namespaces is the MIRIAM Registry and is central in the creation of resolvable Identifiers.org URIs.

3 MIRIAM Registry

The MIRIAM Registry (*see* **Note 15**) is a product of the MIRIAM Guidelines. Having identified the need for adding metadata to model files, it was necessary to create a suitable repository of approved namespaces: MIRIAM Registry. Importantly, the way in which information is structured in this registry takes into consideration the way information is presented and distributed in the scientific domain (Fig. 1).

For the purposes of simplifying data access and information storage within the Registry, an abstraction is made of a "pool" or "set of data" of interest and is referred to as a "data collection." Each data collection is assigned a namespace, which is human-readable ("taxonomy," Fig. 1). This data collection contains a finite number of data records, each of which exists in this namespace regardless of where the data itself is located. Hence, neither data collection nor individual records are restricted by geolocation or database architecture and are thought to exist as abstract concepts. Each record is of course assigned an identifier

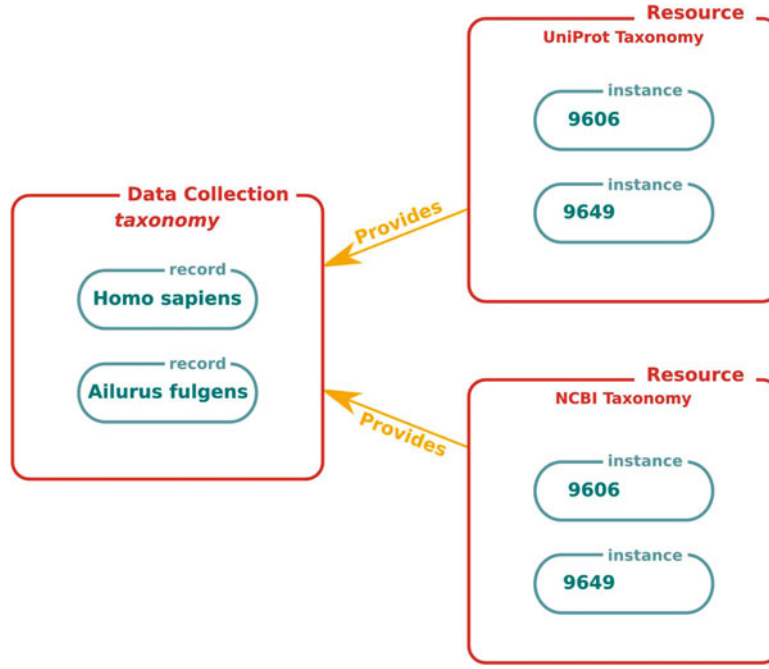
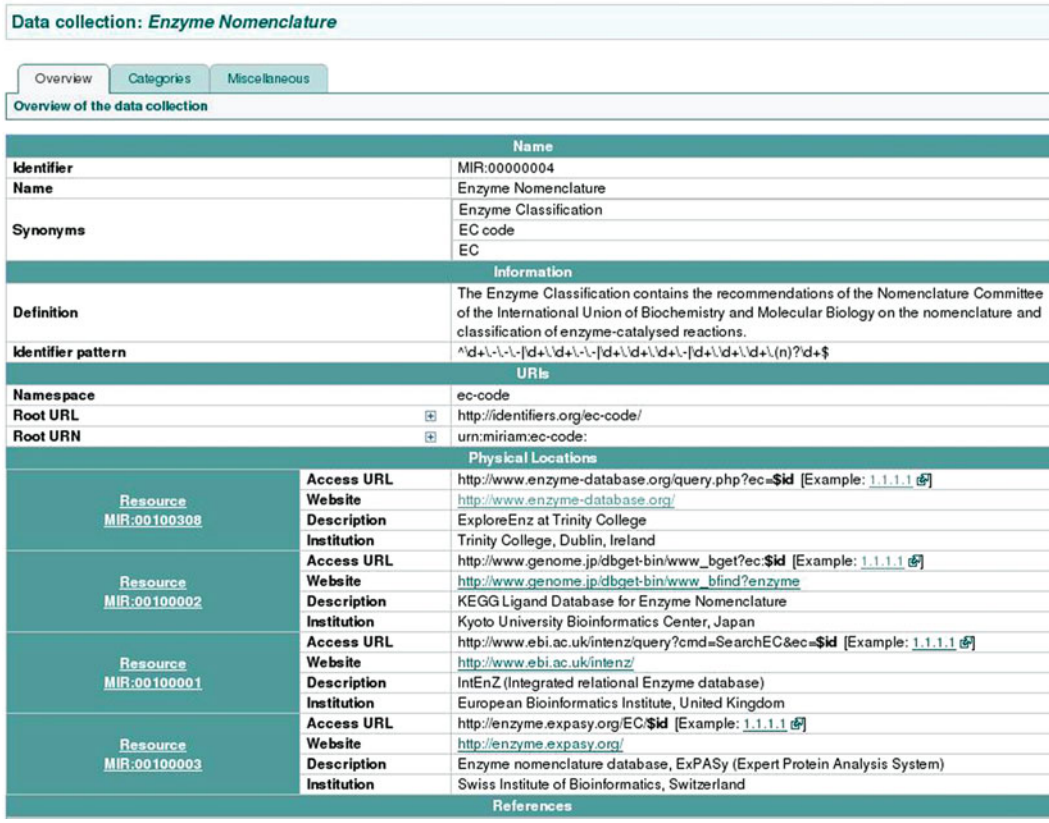


Fig. 1 Structure and nomenclature of information stored in the MIRIAM Registry

by the data providers themselves. For each data collection, there may be one or more "resources" that serve the pertaining information. These "resources," then, are the physical locations where the data itself may actually be accessed, if required. In the example shown (Fig. 1), both the "UniProt" and "NCBI" *resources* provide access to *instances* of *records* from the "Taxonomy" *data collection*.

This simple separation of the data (record) from the locations where the information can be accessed (resources) allows the building of a robust, unambiguous, and perennial identification and cross-referencing system.

The namespace information stored in the Registry can then be used to construct URIs of either URN or URL forms. This requires, besides the namespace assigned and stored in the Registry, a unique collection-specific identifier (generated by the data provider). Since the same namespace is used in both URN and URL forms, and the identifier for a particular record is fixed, it is apparent that both forms are highly related, and indeed it is possible to convert from one form to the other. A typical Registry entry is provided (Enzyme Nomenclature, Fig. 2). It should be stated that the Identifiers.org URLs (cf. below) are the preferred form of identifiers, and that the URN form is becoming largely deprecated, given all the advantages the URL form presents.



Data collection: *Enzyme Nomenclature*

Overview | Categories | Miscellaneous

Overview of the data collection

Name		
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		

Fig. 2 MIRIAM Registry entry for the enzyme nomenclature data collection

A brief description of the variety of information captured for each data collection in the Registry, and its significance, is given in Table 3.

3.1 Registry Accessibility Features and Facilities

A variety of user-centric features have been provided alongside the Registry to facilitate both its efficient use and to encourage its rapid adoption. These include Web Services to allow programmatic access to the Registry [6,7], for example to validate, resolve, or create MIRIAM URIs.

Other useful features:

Collection "*tags*": A few tags, taken from a defined set of keywords, are associated with each data collection. They describe either the type of information recorded by the collection ("sequence," "phenotype"), the subject of that collection ("gene," "drug"), the domain area to which it relates ("disease," "neuroscience"), or the taxonomic relation of the data ("mammalian," "human"). This allows users to identify collections of interest. The refinements planned for the system are discussed elsewhere [8].

Table 3
Description of the main information components stored for data collections in the MIRIAM Registry

Information field	Description	Significance/comment
Collection name	A human-readable name to refer to the collection	Usually assigned based upon the most commonly associated resource for the collection
Collection identifier	A unique identifier for the collection within the Registry	Not intended for human readability
Collection synonyms	Other names by which the collection may identified	This field is searchable through the web interface, and query-able through web services
Collection identifier pattern	A regular expression pattern that matches all valid identifiers within the collection	Can be used through web services to validate potential identifiers
Collection namespace	The namespace assigned to the data collection	Can be used to construct URIs, and is usually an acronym based upon the collection name, or based upon the most commonly associated resource(s)
Access URLs (resources)	The physical location URL which can be used to access to a given record from the associated collection	URLs can be modified if needed by Registry curators, allowing its seamless use to the community Each URL is attached to a resource (which is also uniquely identified)
References	Reference information for the data collection	Directs to citation information, or user guides

Resource health: A "health status" has been implemented at the level of the individual resources listed in the Registry, whereby a daily health check is automatically performed for each resource. This is summarized on the data collection listing for each resource, where it is depicted by color coding of the "Resource identifier" panel. A calendar view of the uptime and further details are also available. This system is also used by the Registry curators to identify issues with resources.

Registry download: The entire contents of the Registry can be downloaded in XML format, through the "Export" link on the left panel of any MIRIAM Registry page. This is often preferred by users who would otherwise need to perform numerous queries through Web Services.

Submission of new data collections: The Registry aims to provide its services to any domain of the biological sciences. Any users wishing to submit a collection for inclusion can use the "Submit new" feature. In addition, anyone can provide suggestions for modifications/improvements to the presented information. As a community-driven project, we welcome and encourage all such submissions.

Search facility: It is possible to search the information stored in the Registry using the provided search functionality. This search functions against all textual information stored, such as the collection name, synonyms, and collection description.

Flag system: Over the course of time, the Registry has evolved from collating only "free-to-use" collections and resources, to now accommodating those which may not be free, or have other restrictions. Of course it is useful to present information on such restrictions to the users, since it can affect their choice of collection or resource to use. This is achieved through the use of a "flag" system. Current flags include, for example, "License restriction" (which may preclude access or use for commercial purposes), and "Access restriction" (e.g., requiring registration). Data collections with restrictions are clearly labeled.

3.2 MIRIAM URIs

As stated, the namespace stored in the MIRIAM Registry can be used to construct both URN and URL forms of identifiers. While initially URNs were recommended for use in annotation, an increasing number of users expressed the desire to process these annotations in situ. For instance, given an identifier for a chemical compound in the ChEBI (*see* **Note 16**) data collection, it may be desirable to know if this model component is identical to a component in another model that was annotated with a PubChem (*see* **Note 17**) collection-based annotation. Using the URN form one would need to, for example, perform queries via web services to retrieve resolving locations (resources) for that URN, then to examine any common cross-references and descriptions contained on each target page (one for each CheBI and PubChem record). The provision of URL-based identifiers removes one step in this process, and depending on how such an information was retrieved, provides additional information in various formats (such as RDF/XML).

MIRIAM URIs are composed of four parts partitioned by a separator, "/" for URLs and ":" for URNs. The stem of the construct is constant and is composed of the definition of URI form, e.g., http:/, and the definition of the URI type, e.g., identifiers.org. The next part specifies the data collection to be identified, using the namespace recorded in the MIRIAM Registry, for example pubmed. Finally, the record identifier, which is unique and assigned by the data provider, for example 16333295, is appended to construct the full identifiers.org URL, http://identifiers.org/pubmed/16333295.

3.3 Identifiers.org URLs

Identifiers.org (http://identifiers.org, *see* ref. ([8]) is a resolving layer built upon the information stored in the MIRIAM Registry and provides resolvable identifiers. Each collection in the Registry has an associated namespace and dictates the syntactic stem that is to be used to construct both URN and URL forms of identifiers.

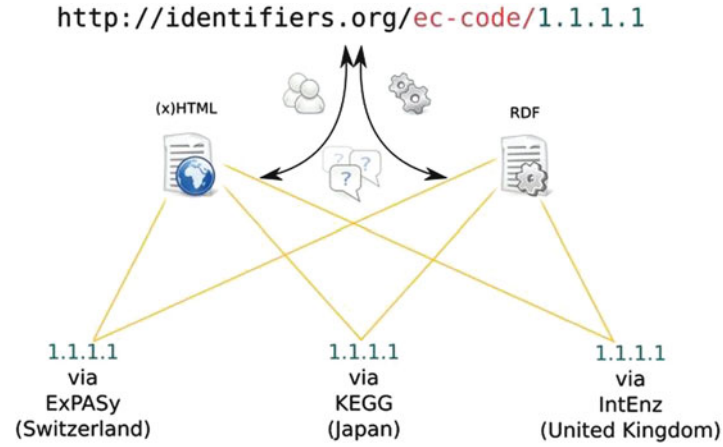


Fig. 3 Illustration of the relationship between the intermediate resolving location and physical locations associated with a specific data collection

A collection record can then be specified using the collection-specific identifier assigned by the data provider. For comparison, both URN (top) and URL (bottom) forms identifying the MIRIAM publication in the PubMed data collection are shown:

urn:miriam:pubmed:16333295

http://identifiers.org/pubmed/16333295

Since the URL above specifies a record, it may be associated with any number of resolving locations (resources). Hence, since it is preferable to provide all of them rather than preselecting a single one, the URL instead resolves to an intermediate page, where all such locations are presented to the user for selection. This behavior is depicted in the illustration below (Fig. 3).

The Identifiers.org URL form also allows various levels of customization in resolving behavior, for example allowing one to request the resolved information to be returned in a specified format, such as RDF.

3.4 Identifiers.org Granularity

Identifiers.org URLs can be used directly to access the information available in the MIRIAM Registry. The following examples illustrate how to build URLs at an appropriate level of granularity.

Identification of a data collection.

The URL below resolves to the entry for the "PubMed" collection in the MIRIAM Registry. http://identifiers.org/pubmed/: Since MIRIAM itself is a collection (of namespaces) and is listed in the MIRIAM Registry, it is also possible to reference the "PubMed" collection using the identifier for "PubMed" in the MIRIAM Registry (MIR:00000015), allowing retrieval of the same entry in the database with the synonymous URL: http://identifiers.org/miriam.collection/MIR:00000015

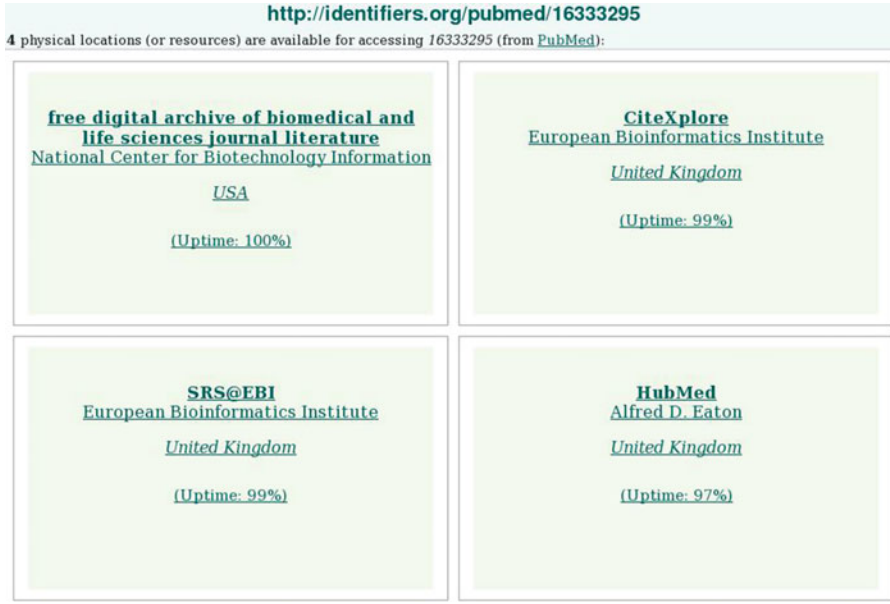


Fig. 4 Illustration of an example intermediate page which is displayed when accessing a "PubMed" data record

Identification of a record within the PubMed data collection.

The URL below resolves to an intermediate page which lists all available resolving locations listed for this collection, in the MIRIAM Registry.

http://identifiers.org/pubmed/16333295: The intermediate page corresponding to this example is shown (Fig. 4).

For convenience, listed alongside each associated resource is its name, geographical location, and its "uptime," summarized from the health check status.

Of course, a user will likely have, or develop over time, a preference for one resource over another, for whatever reason. In such instances, they may wish to directly and repeatedly resolve to that specific resource location. This can be accomplished using the resource identifier associated with each collection resource.

http://identifiers.org/pubmed/16333295?resource=MIR:00100023: In this case, the page corresponds to http://www.ncbi.nlm.nih.gov/pubmed/16333295, which is the NCBI resource location associated with the PubMed data collection (Fig. 5).

Of course, it is not convenient or likely that users will commit individual resource identifiers to memory. This issue necessitated the creation of the "profile" parameter.

3.5 Profiles

Profiles allow one to customize the behavior of the resolving system through pre-selection of the resources to be used in dereferencing Identifiers.org URLs. This means that one can define a set of

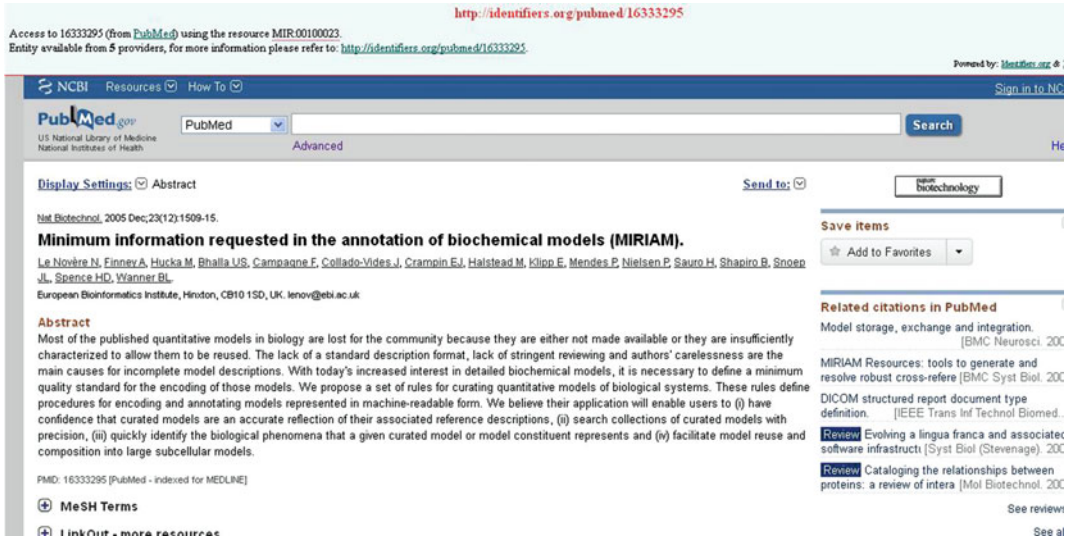


Fig. 5 Accessing data through a specified resource, using an Identifiers.org URI

resolving locations for, potentially, every data collection stored in the Registry. Currently the number of available profiles is limited to those created by the Registry curators. Work is under way to extend this facility and allow users to create and share their own profiles. Profiles will be allowed to be "private," while an interface is being implemented to allow public profiles to be searched. For example, the predefined profile "most_reliable" (below) always returns the instance of a record through the resource with the highest uptime. The "most_reliable" profile is based on the health check history of the resource. http://identifiers.org/pubmed/16333295?profile= most_reliable

The use of the URL form, combined with judicious "parameters," simplifies access to a wealth of information, largely obviating the need for directly querying the Registry through web services. However, it should be noted that the use of the URL form for identification purposes should not incorporate the use of any parameter. Hence, in the unambiguous and perennial identification of data, the identifier should be considered as being the minimal string that specifies a record. From a practical perspective, those users who have in the past used identifiers of the URN form can convert them into identifiers.org URLs if they choose, or indeed vice versa.

3.6 BioModels.net Qualifiers

The purpose of "Qualifiers" is to refine the relationship between, for example, a model component and the resolved target of a cross-reference associated with that component. In the absence of a qualifier, the relationship assumed is an "is" relationship. For instance, given a model component labeled as "Glu" containing

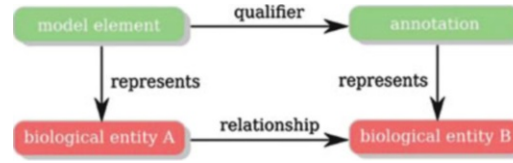


Fig. 6 Schematic representation between model component in a file (model element), the "real-life" entity it seeks to represent (biological entity A), the external resource annotation provided with it (annotation), and the "real-life" target of that external resource annotation (biological entity B)

an unqualified annotation (http://identifiers.org/obo.chebi/CHEBI:17234, which resolves to a page with information about "Glucose"), it should be assumed that the model component (written as "Glu") "is" "Glucose" (the external resource record representing the real-life glucose molecule).

The "is" or "identity" relationship is straightforward to understand, but other qualifiers exist to express more complex relations between model component and an external resource. It should be noted that there are two types of qualifiers, *biological* (in the "bqbiol" namespace) and *modeling* (in the "bqmodel" namespace), which relate either biological/physical objects (genes, proteins, enzymes) or modeling objects/concepts (model files, databases, literature), to model components. Figure 6 illustrates the relationship between model component in a file (model element), the "real-life" entity it seeks to represent (biological entity A), the external resource annotation provided with it (annotation), and the "real-life" target of that external resource annotation (biological entity B).

For example, expanding on the example above, a model of glycolysis may contain a model component labeled "PFK" (model element), representing the "real-life" enzyme phosphofructokinase (biological entity A). The external resource annotation presented alongside it (annotation), when resolved, can be used to represent a database record for the real-life activity of the phosphofructokinase enzyme (biological entity B), which is important with respect to its function in the model, namely its catalysis of a specific reaction. In this case, an appropriate qualifier would be "hasProperty." The qualifier is, in essence, a reflection of the relationship between two representations, one being held in a model, and the other in an external resource. This is necessary since it is not possible to actually attach a PFK molecule to an electronic file, whether it is a model file, or a database record.

Some of the biological relationships that can be represented are shown below, with reference to the figure above. It should be noted that each qualifier is presented in two forms, noun and verb, to allow users to select whichever they are most comfortable with. Both can be used synonymously. The full list of qualifiers is available

Table 4
Examples of the biological qualifiers available to refine the relationships between model component and external resource

Qualifier	Description
bqbiol:hasPart bqbiol:part	The biological entity represented by the model element includes the subject of the referenced resource (biological entity B), either physically or logically. This relation might be used to link complex to the description of its components
bqbiol:isDescribedBy bqbiol:description	The biological entity represented by the model element is described by the subject of the referenced resource (biological entity B). This relation should be used, for instance, to link a species or a parameter to the literature that describes the concentration of that species or the value of that parameter
bqbiol:isEncodedBy bqbiol:encoder	The biological entity represented by the model element is encoded, directly, or transitivity, by the subject of the referenced resource (biological entity B). This relation may be used to express, for example, that a protein is encoded by a specific DNA sequence
bqbiol:isHomologTo bqbiol:homolog	The biological entity represented by the model element is homologous to the subject of the referenced resource (biological entity B). This relation can be used to represent biological entities that share a common ancestor
bqbiol:occrsIn bqbiol:container	The biological entity represented by the model element is physically limited to a location, which is the subject of the referenced resource (biological entity B). This relation may be used to ascribe a compartmental location, within which a reaction takes place
bqbiol:isVersionOf bqbiol:hypernym	The biological entity represented by the model element is a version or an instance of the subject of the referenced resource (biological entity B). This relation may be used to represent, for example, the "superclass" or "parent" form of a particular biological entity

from the BioModels.net website (*see* **Note 18**) and can be expanded and refined upon community request and feedback (Table 4).

3.7 Incorporating Qualifier Relationships

The simplest way to understand qualifiers is to consider them as being the "predicate" in a "subject, object, predicate" sentence, where the subject is the model component, the object is the target of the external resource annotation, and the predicate is the qualifier relationship between them.

Qualifiers are commonly used within metadata in model encoding formats, such as SBML:

1. <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
2. xmlns:bqbiol="http://biomodels.net/biology-qualifiers/">
3. <rdf:Description rdf:about="#MyModelElement">
4. <bqbiol:hasPart>
5. <rdf:Bag>

```
6. <rdf:li rdf:resource="http://identifiers.org/uniprot/P04551"/>
7. <rdf:li rdf:resource="http://identifiers.org/uniprot/P10815"/>
8. </rdf:Bag>
9. </bqbiol:hasPart>
10. </rdf:Description>
11. </rdf:RDF>
```

The use of Identifiers.org URLs does not in itself require any particular format or syntax. It can therefore be incorporated into any structured format relatively easily. However, many structured formats do themselves have a syntactic procedure through which such annotations are to be expressed. For example, within SBML, such annotations are encoded in RDF (*see* **Note 19**) blocks.

A detailed line-by-line description of the example above:

1. The <rdf element open tag and definition of XML namespace declaration for RDF use.
2. Definition of the biology-qualifiers namespace.
3. RDF Description block opened, with the subject being MyModelElement
4. The qualifier for the block is hasPart from the bqbiol namespace.
5. The rdf:Bag construct allows the inclusion of multiple URIs in an annotation.
6. The li line element where a resource is specified.
7. The li line element where a resource is specified.
8. Close tag to end the Bag block.
9. Close tag to end the hasPart block.
10. Close tag to end the Description block.
11. Close tag to end the RDF block.

This annotation block should be interpreted to mean that "MyModelElement" represents a biological object that "has parts" described by the records in the UniProt data collection specified by the identifiers P04551and P10815. In this example, the UniProt specified entries refer to Cyclin-dependent kinase and G2/mitotic-specific cyclin cdc13, which are both involved in the control of the cell cycle at the G2/M (mitosis) transition.

3.8 Alternative Identification Schemes

The Identifiers.org identification scheme offers distinct advantages over some other well-known systems, some of which are described briefly below.

Persistent Uniform Resource Locations (PURLs) (*see* **Note 20**) are subtly different in intent from Identifiers.org URLs. Since this is an open system, in the sense that once registered any individual may create a PURL, there can potentially be a plethora of different PURLs that all identify the same record (have a common endpoint). This is an hindrance to data integration. In addition, the focus of PURLs is to permanently identify a record resolved through a specified resource, thus effectively tying a record identifier to a specific instance, within a single URL. This should be contrasted with a record identifier using Identifiers.org URLs, which can be used to resolve information through any number of associated resources.

Digital object identifiers (DOIs) (*see* **Note 21**) are generally associated with online authored publications, and hence may not be as well suited to the referencing of biological entities. In addition, it is a fee-based assignment service, and the identifier designated by DOI does not reuse the identifier assigned by the data provider. Finally, like PURLs, a DOI resolves to a single instance of a record.

Life Science Record Names (LSRN) (*see* **Note 22**) are the closest relative of the MIRIAM identification scheme, in that they use a central database with assigned namespaces and store information on the associated resolving locations. The key differences lie in the extensive curation of the MIRIAM Registry, its broader coverage, and the supporting facilities it offers, including web services, programmatic access to the database, health check, XML download availability, together with an extensive and highly active community of users.

A more complete comparison of these, and other, identification schemes is espoused on the Registry website (*see* **Note 23**). There follows a summary of the key advantages proffered by the MIRIAM system:

- Open submission—Anyone can make a submission to the Registry.
- Curated—The content of the Registry is heavily curated and maintained for accuracy by a dedicated curation team.
- Resolution system—The scheme adopted allows the mapping of records to multiple resolving locations.
- Health check—Daily monitoring of all resources, with curator intervention when necessary.
- Extensive support—A growing community of users to provide software and tools in support of the system (see below).
- Accessibility—A variety of access methods is provided, including web services.
- Export—The entire content of the can be exported as XML, allowing noninteractive processing of Registry content.

- Free—There is no restriction on the use of information in the Registry, and no registration requirement.
- Standardization—MIRIAM is itself a partner in a number of standardization efforts.

3.9 The Registry User Community

There are a number of perspectives that can be taken on the knowledge captured in the MIRIAM Registry. It can be viewed as part of a standardization effort, thus having associated compliant file formats, and supporting software and tools; it can be regarded as a way to identify both data records and a means to standardize namespaces and associated resources; it can also be considered within the landscape of other, sometimes competing, identification schemes. Each of these perspectives is briefly addressed below.

Since many structured formats conform to the MIRIAM Guidelines, they by default should use annotations based on the MIRIAM Registry. Since SBML is one such structured format, all tools that read, write, or manipulate this format will intrinsically handle Identifiers.org URIs. This covers a broad spectrum of activities ranging from the annotation of models, through processing of those models to do novel research, to creating human or machine-readable representations of those models.

Identifiers based on information stored in the Registry are already widely used, most notably within BioModels Database (*see* **Note 24**, ref. [9]). The latest release of the database (22nd release, May 2012) contains over 142,900 models, with over 444,130,000 annotations. These model files are available freely and can be downloaded with either URN or URL annotations, with the latter being the default annotation style.

Since the Registry also assigns and stores namespace information for data collections, as well as associated synonyms, this knowledge itself can also be used to harmonies or standardize resource nomenclature. For example, both the PSI-MI (Proteomics Standards Initiative—Molecular Interactions; *see* ref [10]) and BioPAX (Biological Pathway Exchange; [11]) working groups use this information to assign standard database names in their controlled vocabularies, using the stored synonym information.

As a standardization effort, support for Identifiers.org URIs, and particularly the use of resolvable identifiers, is growing; LSRN, has announced that it will be transitioning its information into the Registry and will cease further support and development of its own identification scheme. This process is already well under way.

Further information on the formats, tools, and software that utilize MIRIAM Registry information is available from the Registry's documentation (*see* **Note 25**).

3.10 Future Perspectives

The MIRIAM Registry is a stable resource which provides an identifier scheme, a perennial URI generation service, and a resolving system. While its foundations lie in Computational Systems Biology, it is by no means restricted to that domain, and indeed data collections from more diverse fields are continually being incorporated. This potential for its use as a universal cross-referencing system, which was noted during the inception of the system, is now being realized. There should be no impediment in its use in any domain.

The user interface and access options are being continually improved, permitting not only the creation of perennial and unambiguous identifiers, but also facilitating the customization of the way the underlying data is addressed. The ability to create "Profiles," for example, will allow the creation of entire sets of resolving preferences, which can potentially be shared at an institutional, community, or group level.

There previously existed various restrictions governing the suitability for inclusion of data collections into the Registry. These have recently been removed in recognition of the referencing needs of the user community at large. For instance, some proprietary data collections were deemed unsuitable since they required either registration or were subject to fee-based access. The provision of the "flag" system discussed above has enabled the incorporation of such historically non-compliant data sources.

When deliberating upon the future of data access on the web, one must also consider the importance of efforts such as the Semantic Web (*see* **Note 26**) and the Linking Open Data (LOD) (*see* **Note 27**) initiative. In providing resolvable URIs the Identifiers.org addresses some of the demands of this growing community.

The MIRIAM efforts (Guidelines, Registry, and Identifiers.org) are all partners in larger community level standardization efforts, such as MIBBI and BioDBCore [12], as well as members of the modeling community, particularly through their involvement in SBML, but also within the COMBINE (*see* **Note 28**) community.

4 Notes

1. http://www.w3.org/TR/REC-xml/
2. http://sbml.org/Documents/Specifications
3. http://obofoundry.org/
4. http://www.ebi.ac.uk/sbo/
5. http://biomodels.net/kisao/
6. http://mibbi.org/
7. http://www.ncbi.nlm.nih.gov/taxonomy
8. http://www.boldsystems.org/views/taxbrowser_root.php
9. http://www.ncbi.nlm.nih.gov/pubmed

10. http://www.ebi.ac.uk/ena
11. http://www.uniprot.org/
12. http://tools.ietf.org/html/rfc3986
13. http://en.wikipedia.org/wiki/Uniform_Resource_Name
14. http://www.tcdb.org/
15. http://www.ebi.ac.uk/miriam/
16. http://www.ebi.ac.uk/chebi/
17. http://www.ncbi.nlm.nih.gov/pccompound
18. http://biomodels.net/qualifiers/
19. http://www.w3.org/TR/REC-rdf-syntax/
20. http://www.purl.org/
21. http://www.doi.org/
22. http://lsrn.org/
23. http://www.ebi.ac.uk/miriam/main/mdb?section=uris
24. http://www.ebi.ac.uk/biomodels/
25. http://www.ebi.ac.uk/miriam/main/mdb?section=use
26. http://www.w3.org/2001/sw/
27. http://linkeddata.org/
28. http://co.mbine.org/

5 Funding

This work has been supported by EMBL, the ELIXIR Preparatory Phase Project, and by BBSRC grants.

References

1. Hucka M et al (2003) The Systems Biology Markup Language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19:524–531
2. Smith B et al (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25: 1251–1255
3. Courtot M et al (2011) Controlled vocabularies and semantics in systems biology. Mol Syst Biol 7:543
4. Taylor C et al (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol 26:889–896
5. Le Novère N et al (2005) Minimum Information Requested in the Annotation of biochemical Models (MIRIAM). Nat Biotechnol 23: 1509–1515
6. Laibe C, Le Novère N (2007) MIRIAM Resources: tools to generate and resolve robust cross-references in Systems Biology. BMC Syst Biol 1:58
7. Li C et al (2010) BioModels.net Web Services, a free and integrated toolkit for computational modelling software. Brief Bioinform 11: 270–277
8. Juty N et al (2012) Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. Nucleic Acids Res 40: D580–D586
9. Li C et al (2010) BioModels Database: an enhanced, curated and annotated resource for published quantitative kinetic models. BMC Systems Biol 4:92
10. Orchard S, Hermjakob H (2008) The HUPO proteomics standards initiative–easing communication and minimizing data loss in a changing world. Brief Bioinform 9:166–173
11. Demir E et al (2010) The BioPAX community standard for pathway data sharing. Nat Biotechnol 28:935–942
12. Gaudet P et al (2011) Towards BioDBcore: a community-defined information specification for biological databases. Nucleic Acids Res 39: D7–D10