DATABASE
The Journal of Biological Databases and Curation

Database, Vol. 2013, Article ID bat017, doi:10.1093/database/bat017

Original article

Towards the Collaborative Curation of the Registry underlying identifiers.org

Nick Juty[1], Nicolas Le Novère[2], Henning Hermjakob[1] and Camille Laibe[1,]*

[1]Proteomics Services, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [2]Babraham Institute, Babraham Research Campus, Cambridge CB22 3AT, UK

***Corresponding author:** Tel: +44 (0)1223 494 444; Fax: +44 (0)1223 494 468, Email: laibe@ebi.ac.uk

Submitted 30 November 2012; Revised 1 March 2013; Accepted 8 March 2013

Citation details: Juty,N., Novère,N.L., Hermjakob,H., *et al.* Towards the collaborative curation of the registry underlying identifiers.org. *Database* (2013) Vol. 2013: article ID bat017; doi: 10.1093/database/bat017

The MIRIAM Registry (http://www.ebi.ac.uk/miriam/) records information about collections of data in the life sciences, as well as where it can be obtained. This information is used, in combination with the resolving infrastructure of Identifiers.org (http://identifiers.org/), to generate globally unique identifiers, in the form of Uniform Resource Identifier. These identifiers are now widely used to provide perennial cross-references and annotations. The growing demand for these identifiers results in a significant increase in curational efforts to maintain the underlying registry. This requires the design and implementation of an economically viable and sustainable solution able to cope with such expansion. We briefly describe the Registry, the current curation duties entailed, and our plans to extend and distribute this workload through collaborative and community efforts.

Introduction

The annotation of data has become increasingly important, particularly with the advent of high throughput data generation from large-scale 'omics' initiatives. This change in data generating methodologies has decreased the extent of direct human interaction with data sets, simply due to their immense size and complexity. Instead, there is growing reliance on automated or semi-automated computational (pre-)processing. It is in this computational processing that metadata plays a major role, facilitating comparison or integration across seemingly, to the human eye, disparate data.

Most data providers identify individual records within their data sets using alphanumeric identifiers. However, this alone is often not enough to uniquely identify the entity under consideration, especially where multiple data sets are considered simultaneously. Such situations are commonly encountered when processing data from divergent sources. For example, simply considering only taxonomic resources, one can find vastly different species, which use the same identifier: '9606' identifies Homo sapiens (human) in the NCBI taxonomy (1), Bombycilla cedrorum (bird) in the Barcode of Life Data System (2), and Catha edulis (plant) in the Germplasm Resources Information Network Plant Taxonomy (3). It is therefore clear that to generate globally unique identifiers (which allow processing across multiple data sources), identifiers local to data sets are not enough, and more contextual information is needed.

Besides the identification of data entities, another important consideration is the ability to contend with the ever-changing landscape of databases: they can be deprecated or removed over time, often reflecting the sustainability of the funding that drives them or, alternatively, the affiliation of the hosting body may change. These events may lead to changes in the way the data can be accessed. From a practical point, this will often involve changes in the physical access locations (or Uniform Resource Locators; URLs).

MIRIAM registry

The same requirements for an identification scheme existed within the systems biology domain to allow global identification and annotation of model entities across models encoded in the Systems Biology Markup Language (4). This led to the creation of a generic identification scheme, relying on the Uniform Resource Identifier (URI) standard.

To build unique URIs, the system combines two pieces of information: a 'namespace' and the local identifier. The namespace allows the identification of the 'data collection' (a set of data based on some specific perspective given by the data generator) from which the entity to be identified comes from. The local identifier is provided by the original data generator of the collection.

To achieve this, the existence of a centralised list of data collections (and their associated namespace) was required. This led to the launch of the MIRIAM Registry (5) (http://www.ebi.ac.uk/miriam/) to fulfil this need. Each collection listed in the Registry is associated with a namespace and 'resources' through which the primary data can be accessed and retrieved on the web (by means of URLs). Additional information is also stored with each record, such as an identifier pattern (used to perform validation checks during the generation/resolving processes), publications describing the collection, keywords (which can be used for example to find collections about 'protein' or 'pathway') and so forth.

To provide directly dereferenceable URIs and comply with the second rule of Linked Data (6) and be directly usable by Semantic Web applications, the generated URIs are actually HTTP URIs (or Uniform Resource Locators). The resolving services are provided by Identifiers.org (7) (http://identifiers.org/).

The rationale behind the creation of the Identifiers.org URI scheme is to provide globally unique, perennial and standard compliant identifiers for use in a variety of scenarios (from sharing simple web links to storing robust cross-references in a database). In addition, these URIs allow one to access the identified data (or information about the data) on the web, can easily be generated, and provide a plethora of benefits: they do not require foreknowledge of a data provider current, prior or future access URL(s), nor do they necessitate knowledge of its institution or its affiliation to standardisation bodies (such as the OBO Foundry for ontologies) and so forth.

A project, hosted on SourceForge, provides ways to contribute to both the content and the infrastructure: http://sf.net/p/identifiers-org/.

By generating free, unique, perennial and location independent identifiers, Identifiers.org URIs and the underlying Registry have proven to be highly successful and are now widely used outside the modelling field. Numerous 'collections' (ontologies, databases, etc.,) are already using these URIs, including BioModels Database (8), Reactome (9) and SABIO-RK (10). More recent adopters include OpenPHACTS (11) (in the mapping of chemical identifiers from different datasets), Bio2RDF (12) (which provides identifiers.org URIs in their Resource Description Format (RDF) data), PSICQUIC (13) (in their BioPAX export) or the European Bioinformatics Institute (when providing data sets in RDF).

Registry content and maintenance

The Registry is a fundamental component of the whole system: the ability to create URIs relies on the record of the relevant data collection. It was therefore crucial that the Registry can be used freely by the community and that anybody should be able to request the creation of new records.

In fact, besides accepting requests for new collections (via the 'Submit new' feature), the Registry allows the update and improvement of existing collections (via the 'Suggest modifications to this data collection' functionality). Neither page requires extensive data entry, with many of the fields being optional.

However, to insure a high level of consistency and quality between all the information it stores, the Registry is a curated resource. This means that a curator (a person tasked with manually managing data from submission through to its public release) will always review the information provided (whether it is a new submission or a suggested update) and decide whether it is suitable for immediate incorporation. Therefore, all publicly submitted information is verified as being complete, correct and, importantly, up-to-date.

Currently, as part of the curation process for a new submission, a curator will assign a namespace, usually based on the acronym by which the data provider or generator commonly refer to their data set. Where a data set can be divided into different components, for example Kyoto Encyclopedia of Genes and Genomes (KEGG) being divisible into 'Drug', 'Pathway' and 'Reaction' data sets, the namespaces assigned will reflect this hierarchy. Namespaces are usually composed of at least three alphanumeric characters (including 'dash' characters), with a 'period' used to designate subclass relationships (when a data provider actually gives access to multiple data sets). Namespace clashes are not possible, as each is assigned only once and never reused (even in the exceptional case where deprecation is required).

The Registry records contain a brief description of the data collection contents, often gleaned from the data provider themselves. The physical locations where information for a data collection can be retrieved are recorded, along with a functional example identifier, which allows their concatenation into a resolvable URL address. This is actually used in an automated process, which checks resolvability (using the recorded example identifier) on a daily basis. For this reason, during curation, a few keywords from the resolved URL are stored alongside, to enable the detection of URLs that do not resolve properly (whether because the server is down or the URL to access the data has changed).

In this way, a history for the individual resources serving a collection can be collected, allowing the user to select those which they deem most reliable.

Synonyms are also stored for individual data collections and can be used to normalize data annotations, which use legacy identification schemes, specially when this data comes from different sources. For example, such harmonization is undertaken by Pathway Commons (14) when the system processes BioPAX (15) files coming from multiple sources.

Supplementary information, such as documentation related to the collection, is also recorded, along with the institutional details of the data provider.

Following the addition of this core set of information, each collection is further associated with a few 'tags'. These were created by Registry's curators in *ad hoc* fashion when the system was initially implemented, to describe, in coarse grain fashion, the content of the collection to which they are ascribed. The current list of tags makes up a small controlled vocabulary, which is used during curation to associate each collection with two to three tags. These depict, for example, the particular domain of a collection, such as 'protein', 'pathway', 'genome' or 'ontology'. In this way, users may search the cloud of tags to discover appropriate data collections, which can be used for their annotations or cross-references. This set of tags will be mapped to other controlled vocabulary terms, such as the Biomedical Resource Ontology (16), to facilitate interoperability.

If the curation process following a user request (whether new submission or update) is successful, the affected data collection is released to the live database immediately. However, there are situations where this is not possible, and the entry is placed into the (private) curation pipeline. Reasons for this include there not being sufficient information to determine what the coverage of the data set is, or if there are discrepancies between the documentation and the data that are live (e.g. identifiers that do not match documented patterns). In these instances, the curator contacts the data provider and will try to resolve any issues encountered.

Recently, we have introduced 'restriction' flags to help users make an informed decision about the usability and suitability (depending on the users' situation and needs) of data collections. For instance, a 'License restriction' is placed on data collections that are not free to all users, and may preclude their use by commercial institutions, and a 'Collection maintenance' flag is associated with data collections not actively maintained. All collections listed with restrictions provide a brief summary of the issue (often with links to the data provider's documentation). A full list of the current 'restrictions' is provided in the Frequently Asked Questions' page of the Registry.

The curation of the Registry is a continuous process, which besides the addition of new collections, must maintain those that are pre-existing. As the primary function of the Registry is to provide stable and resolvable URIs for data records, the accuracy of all possible resolvable locations is checked daily. The accompanying information around specific collections in the Registry (documentation, description, synonyms, etc.,) is not checked routinely, as it is regarded as stable. Individual data providers and distributors are frequently solicited to ensure accuracy of Registry records, whereas users are provided a means to feedback suggestions to improve them.

Community involvement

With the growth in the number of records stored in the Registry, comes an increased curational load. This could be eased, in a sustainable way, through moderated collaborative curation. Moderation is required to ensure a consistent level of quality between records, as different contributors may have diverging opinions on the precise information that should be recorded, for example, what constitutes an appropriate namespace or synonym (alternative name for a data collection), or how to abstract an identifier pattern and construct a regular expression to describe it. We envisage that, in advance of this collaborative community effort, we shall have produced a more encompassing curation guide to facilitate this process. Therefore, over time, the level of moderation required should decline.

Curation by data providers

Currently, submissions and updates to the Registry are curated before their public release. This curation work is done exclusively by delegated individuals. This presents several limitations.

First, curation requires dedicated people to gather all the necessary information for new submission, or to check and complete user submissions. Moreover, regular updates are necessary to keep the full list of resources' URLs in working condition: an automatic daily check does help the curators to detect failures, but manual work is still required to determine the cause and therefore the possible solution, on a case by case basis.

Curation frequently requires correspondence with the resources hosting the data, to resolve ambiguities or issues. It is often not easy to identify the most relevant person(s) to contact to discuss such matters, and even then, they may not be the most suitable to either make or act on the outcomes of any discussion. Therefore, these consultations can be time consuming.

Finally, all this curational burden increases as the Registry grows in content.

To address these points, we intend to involve data providers more directly, and earlier, in the curation process and in the subsequent maintenance of their records in the Registry. By giving providers optional 'ownership' of their record, they would be able to assign and modify namespace(s) as well as maintain supplementary information. As most changes happen at the level of the data providers, this should allow the Registry's content to be updated more quickly. These activities could still be overseen by the Registry's curators, to retain a consistent level of quality and stability across the entire resource. Of course, this would require the update of the underlying software infrastructure, the development of several new user interfaces and the training of people to use the tools quickly, accurately and efficiently.

Customization by users

One of the use cases for the generated URIs is for them to be used at all levels of a data management infrastructure: from internal identifiers in the database to their display in user interfaces. This can already be achieved, but some tools and services prefer to direct their users to a given resource instead of the current choice provided by Identifiers.org (in the case when multiple resources are recorded). One example of this usage is BioModels Database, which store all model annotations (with URIs) and also needs to provide hyperlinks in its web interface for them.

A way to achieve this aim is by developing the concept of 'profiles' and allow any users, project or institution to create them. A profile is a subset of the Registry entries, for example only containing the collections used by its creator. For each collection in a profile, the creator can choose one preferred resolving location. So provided with a profile, each URI would directly resolve to a single location, through the preference expressed in the creator's profile. The advantage of user-defined profiles is highlighted by comparing the resolving of one URI using a profile (http://identifiers.org/genecards/ABL1?profile=most_reliable) and without any profile information (http://identifiers.org/genecards/ABL1). Using a profile allows one to directly access an information record without navigating through the intermediate page listing all the possible resolving locations.

We intend to allow all users to create their own profiles. Moreover, other settings will be available for each profile, for example creators will be able to choose whether it should be public (and could be re-used) or private.

Current and future collaboration with community efforts and registries

The Registry participates in the BioDBCore (17) effort, which aims to provide a 'Minimum Information' checklist (18) for database providers. This records elements such as data release frequency, curation policy and listing of standards, formats and terminologies used within the data set; this is largely complimentary to what is stored in the Registry. Once BioDBCore records are made available by data providers, the Registry will 'pull' them, and they will be used as skeletons, ready for further curation. This should lead to a very productive synergy between the Registry and the BioSharing initiative (19) (which will host this information, making use of Identifiers.org URIs for identification purposes), with both efforts providing complementary elements.

The Registry has an ongoing collaboration with the Bio2RDF project, used by the Linked Data community. The aims are to harmonize URIs, facilitate integration and enable cross-resource queries. Work is currently in progress to expand the Registry to cover all the data sets used in Bio2RDF. Also, Bio2RDF already started to use identifiers.org URIs to present 'owl:sameAs' statements to data resources. Analogously, for relevant data sets, Identifiers .org will provide links to Bio2RDF.

Previous collaboration with the Life Science Record Name resulted in the incorporation of all its records before its retirement. The Registry has also used several other sources of information, such as the database abbreviations of Gene Ontology (GO), and the list of databases cross-referenced in UniProtKB. Some data are also periodically drawn from the yearly *Nucleic Acids Research* Database issue. In most cases, the information is pulled from those sources to the Registry by the curators.

As part of the curation process in the Registry, valuable information about the different databases that serve the biological sciences community is recorded. In many cases, this includes the web services provided by those resources. As the BioCatalogue (20) has been designed to provide this information in great detail, we intend to cross-link to this resource instead of recording duplicate web services information.

Therefore, to minimize the curation effort, we intend to streamline the Registry's content and emphasize our focus on URI generation/conversion services. This means contributing to and re-using all relevant existing efforts whenever possible.

Conclusion

The adoption of Identifiers.org URIs is growing significantly and rapidly within the scientific community. In particular, these URIs are widely used by researchers tasked with placing existing data sets in the Linked Open Data cloud (21). This has started to place a burden on the underlying Registry and, more specifically, the team in charge of curating its content.

This is by no mean unique to the Registry: any resource relying on curation activities (22) is facing the same challenges (23, 24). This burden will become acute in the near future, and one means to address this may be by involving the community of data providers and users. Several resources have already moved in this direction, such as Rfam (25), WikiPathways (26) or Gene Wiki (27) and the initial results are positive. This proves that involving the community in the maintenance of the Registry is a viable solution.

By supplying data provider and users easy to use tools to contribute to the content of the Registry, we should fulfil two major goals: providing content that better satisfies users' needs and reducing the curation burden on our team, even with a growing Registry. This route would offer the additional advantages of being both cost-effective and sustainable in the currently harsh socio-economic environment.

Acknowledgements

The authors would like to express our gratitude to the users for submitting new collections and contributing updates, which greatly help us to provide a reliable service to the community.

Funding

Identifiers.org and its associated Registry have been developed with funding from the Innovative Medicines Initiative Joint Undertaking (115191) (Open Pharmacological Space, OpenPHACTS); the Biotechnology and Biological Sciences Research Council (BB/E006248/1); the European Molecular Biology Laboratory and ELIXIR (Preparatory Phase).

Conflict of interest. None declared.

References

1. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
2. Sarkar,I.N. and Trizna,M. (2011) The Barcode of Life Data Portal: bridging the biodiversity informatics divide for DNA barcoding. *PLoS One*, **6**, e14689.
3. USDA, ARS, National Genetic Resources Program. (2008), Germplasm Resources Information Network (GRIN) National Germplasm Resources Laboratory, Beltsville, Maryland. http://www.ars-grin.gov/cgi-bin/npgs/html/index.pl?language=en (20 March 2013, date last accessed).
4. Hucka,M., Finney,A., Sauro,H.M. *et al*. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531.
5. Laibe,C. and Le Novère,N. (2007) MIRIAM Resources: tools to generate and resolve robust cross-references in Systems Biology. *BMC Syst. Biol.*, **1**, 58.
6. Berners-Lee,T. (2006) Linked data, in design issues: architectural and philosophical points. *W3C website*, http://www.w3.org/DesignIssues/LinkedData.html.
7. Juty,N., Le Novère,N. and Laibe,C. (2012) Identifiers.org and MIRIAM registry: community resources to provide persistent identification. *Nucleic Acids Res.*, **40**, D580–D586.
8. Li,C., Donizelli,M., Rodriguez,N. *et al*. (2010) BioModels database: an enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst. Biol.*, **4**, 92.
9. Croft,D., O'Kelly,G., Wu,G. *et al*. (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.
10. Wittig,U., Kania,R., Golebiewski,M. *et al*. (2012) Sabio-rk–database for biochemical reaction kinetics. *Nucleic Acids Res.*, **40**, D790–D796.
11. Williams,A.J., Harland,L., Groth,P. *et al*. (2012) Open PHACTS: semantic interoperability for drug discovery. *Drug Discov. Today*, **17**, 1188–1198.
12. Belleau,F., Nolin,M.-A., Tourigny,N. *et al*. (2008) Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Inform.*, **41**, 706–716.
13. Aranda,B., Blankenburg,H., Kerrien,S. *et al*. (2011) Psicquic and psiscore: accessing and scoring molecular interactions. *Nat. Methods*, **8**, 528–529.
14. Cerami,E.G., Gross,B.E., Demir,E. *et al*. (2011) Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res.*, **39**, D685–D690.
15. Demir,E., Cary,M.P., Paley,S. *et al*. (2010) The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.*, **28**, 935–942.
16. Tenenbaum,J.D., Whetzel,P.L., Anderson,K. *et al*. (2011) The biomedical resource ontology (bro) to enable resource discovery in clinical and translational research. *J. Biomed. Inform.*, **44**, 137–145.
17. Gaudet,P., Bairoch,A., Field,D. *et al*. (2011) Towards BioDBcore: a community-defined information specification for biological databases. *Nucleic Acids Res.*, **39**, D7–D10.
18. Taylor,C.F., Field,D., Sansone,S. *et al*. (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.*, **26**, 889–896.
19. Field,D., Sansone,S., Delong,E.F. *et al*. (2010) Meeting report: BioSharing at ISMB 2010. *Stand. Genomic Sci.*, **3**, 254–258.
20. Bhagat,J., Tanoh,F., Nzuobontane,E. *et al*. (2010) BioCatalogue: a universal catalogue of web services for the life sciences. *Nucleic Acids Res.*, **38**, W689–W694.
21. Halb,W., Raimond,Y. and Hausenblas,M. (2008) Building linked data for both humans and machines. In: *WWW 2008 Workshop: Linked Data on the Web (LDOW2008), Beijing, China.*
22. Sanderson,K. (2011) Bioinformatics: Curation generation. *Nature*, **470**, 295–296.
23. Burge,S., Attwood,T.K., Bateman,A. *et al*. (2012) Biocurators and biocuration: surveying the 21st century challenges. *Database*, **2012**, bar059.
24. Gaudet,P., Arighi,C., Bastian,F. *et al*. (2012) Recent advances in biocuration: meeting report from the fifth International Biocuration Conference. *Database*, **2012**, bas036.
25. Gardner,P.P., Daub,J., Tate,J. *et al*. (2011) Rfam: Wikipedia, clans and the ''decimal'' release. *Nucleic Acids Res.*, **39**, D141–D145.
26. Kelder,T., van Iersel,M.P., Hanspers,K. *et al*. (2012) WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.*, **40**, D1301–D1307.
27. Huss,J.W., Orozco,C., Goodale,J. *et al*. (2008) A gene wiki for community annotation of gene function. *PLoS Biol.*, **6**, e175.