Identifiers.org: integration tool for heterogeneous datasets

Camille Laibe[1], Sarala Wimalaratne[1], Nick Juty[1], Nicolas Le Novère[2], Henning Hermjakob[1]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD

[2]Babraham Institute, Babraham Research Campus, Cambridge, CB22 3AT

Background

Generally, data providers identify individual records within their datasets using an identifier or accession number. Commonly, those identifiers are only unique within the dataset they originate from. For example, ‘9606’ identifies “Homo sapiens” in the NCBI Taxonomy, but also identifies “Catha edulis” in the GRIN Taxonomy for Plants, the publication “Kohlhausen K. Das Offentliche Gesundheitswesen 1976 38(7):424-430” in PubMed, the “parathyroid hormone” in HGNC, “3-Fluorotoluene” in PubChem, etc. Additionally, the data is often distributed via multiple resources, using different URLs to access individual records. For example, the enzyme nomenclature information is available from at least 4 different sources: ExplorEnz (Trinity College, Dublin), KEGG Enzyme Database (Kyoto University Bioinformatics Center), ENZYME (Swiss Institute of Bioinformatics) and IntEnz (EMBL-European Bioinformatics Institute).

In order to address those issues related to the increased proliferation of ambiguous identifiers and non-perennial access URLs, an effort was launched in 2006 to provide a system through which appropriate URIs (Uniform Resource Identifiers) could be generated, based on existing local record identifiers already assigned by the data providers (http://identifiers.org/registry) [1].

A resolving system (http://identifiers.org/) [2], was launched to support requests from the Semantic Web community to provide these as HTTP URIs. These URIs are directly incorporable in datasets and usable by Semantic Web applications. Software tools handling data using those URIs need little work to process them and display them in a meaningful way to the end user (these URIs can actually be used as they stand in web interfaces). Moreover, these URIs are free, and provide unique, perennial and location-independent identifiers.

Here, the underlying Registry acts as the central storage repository for general information on the various datasets (termed *collections*), *namespace* information (unique short string identifying the collections), and lists the *resources* (physical locations from where data records can be retrieved).

The infrastructure is already used very successfully by, for example, the computational modelling community, which requires the ability to perennially record cross-references and links to external data records, despite the ever changing nature of the location of information on the web. It is being continually improved to meet the growing needs of new user communities.

Results

Identfiers.org is now running on multiple servers in 2 redundant data centres in London, which provide more reliable, robust and faster services to the community.

The semantics of the URIs handled by the resolver have been enhanced by having 2 distinct types of URIs: one for identifying the entity concept, and one for identifying the information recorded by the Registry about it. This means that canonical URIs (which uniquely identify data entities and are of the form http://identifiers.org/[namespace]/[entity]) provide more directly usable information for end users (previously a list all possible resources was provided, now one default resource is displayed, with a link to the former page). This default resource is selected using an algorithm which uses various properties such as resource reliability (‘uptime’) and data ownership (designated ‘primary’ for resources maintained directly by the data provider) status of the resources. Registry records of data entities are identified by URIs of the form http://info.identifiers.org/[namespace]/[entity]. Those URIs can be used to identify and retrieve metadata provided by the Registry or access information in various formats, such as RDF/XML (via content negotiation).

With the growing uptake of Linked Data in life sciences, exemplified by projects such as Bio2RDF [3] and the launch of the EBI RDF platform [4], new functionality and improvements to the infrastructure have been made. For example, there are now numerous URIs used to identify equivalent data records. These pose a challenge when integrating across heterogeneous datasets, even if all are encoded in RDF. In order to facilitate such integration, we have extended the services provided by Identifiers.org in order to serve a SPARQL compliant endpoint for URI schemes conversion. This allows the conversion of a URI from one scheme into a URI from a different one. To allow such functionality, the Registry records various URI scheme formats within the Registry, which are used to generate alternative URIs, using the Identifiers.org URIs as a canonical form. This allows the system to handle modification of the entity identifier, such converting http://identifiers.org/go/GO:0006915 into http://purl.obolibrary.org/obo/GO_0006915. Recording the alternative URI scheme formats requires additional curation efforts, but provides accurate results and should prevent any false positive.

With the growth of Linked Data and Semantic Web efforts, it has also become increasingly important not only to know where data may be accessed, but also the forms or formats in which it is available. Many resources now provide their datasets in a variety of formats on top of the standard HTML format. There are users who specifically require data encoded in a particular format, such as RDF/XML, or JSON. To allow direct data access to records encoded in specific formats, the Registry data model has been extended to enable it to record the various formats provided by a resource and how to access them. This information is stored at the level of the individual resources that offer records for a data collection, since each resource may offer different formats for consumption. Additionally, Identifiers.org has been updated to allow access to this information via content negotiation, performed on a user request.

Conclusion

Identifiers.org URIs are now a core element of numerous data management and provision infrastructures, such as the OpenPHACTS project [5] and the EBI RDF Platform. The extension of the information recorded in the underlying Registry, and the services provided, should greatly help in the integration of heterogeneous datasets. These incremental improvements to the infrastructure facilitate data integration independently of the URI scheme that may have been used originally, and allows the system to be more universally suitable for other use cases.

Acknowledgements

This work received support from the BBSRC (BB/K016946/1), IMI (OpenPhacts/115191) and EMBL.

References

1. Laibe, C., Le Novère, N.: MIRIAM Resources: tools to generate and resolve robust cross-references in Systems Biology. BMC Syst. Biol. 1:58 (2007)
2. Juty, N., Le Novère, N., Laibe, C.: Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. Nucleic Acids Res. 40(D1):D580-D586 (2011)
3. Belleau, F., Nolin, M.A., Tourigny, N., Rigault, P., Morissette, J.: Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. J. Biomed. Inform. 41:706–716 (2008)
4. Jupp, S., Malone, J., Bolleman, J., Brandizi, M., Davies, M., Garcia, L., Gaulton, A., Gehant, S., Laibe, C., Redaschi, N., Wimalaratne, S.M., Martin, M., Le Novère, N., Parkinson, H., Birney, E., Jenkinson, A.M.: The EBI RDF platform: linked open data for the life sciences. Bioinformatics 1–2 (2014)
5. Gray, A., Groth, P., Loizou, A.: Applying linked data approaches to pharmacology: Architectural decisions and implementation. Semant. Web. 0:1–13 (2012)