Uniform Resolution of Compact Identifiers for Biomedical Data

Sarala M. Wimalaratne[1 *], Nick Juty[1 *], John Kunze[2 *], Greg Janée[2], Julie A. McMurry[3], Niall Beard[4], Rafael Jimenez[5], Jeffrey Grethe[6], Henning Hermjakob[1] and Tim Clark[7,8]

1. European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire UK.
2. California Digital Library, University of California, Oakland CA USA.
3. Oregon Health and Science University, Portland OR, USA
4. University of Manchester, Manchester UK
5. ELIXIR, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire UK
6. University of California, San Diego
7. Massachusetts General Hospital, Boston MA, USA 8. Harvard Medical School, Boston MA, USA

*These authors contributed equally to the work.
Corresponding author: Tim Clark (twclark@mgh.harvard.edu)

Abstract

Compact identifiers have been widely used in biomedical informatics both formally and informally. They consist of two parts: 1) a unique prefix or namespace indicating the assigning authority and 2) a locally assigned database identifier sometimes called an accession number. The former is used to avoid global identifier collisions when integrating separately managed datasets that are run by different communities and consortia under a variety of autonomous data management systems and practices. This bi-partite identifier approach predates the invention of the Web, but can be leveraged to work more harmoniously with it.

Identifiers.org and N2T.net are two meta-resolvers that take any given identifier from over 500 source databases and reliably redirect it to its original source on the Web. Identifiers.org is based at the European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL-EBI) and serves the biomedical domain; whereas N2T.net (Name-to-Thing) is based at the California Digital Library (CDL), University of California Office of the President, and is domain-agnostic. Both resolvers, while derived from independently developed code bases, with different features and objectives, can now uniformly resolve compact identifiers using a set of common procedures and redirection rules. Here we report on significant further work by our teams toward a more unified approach to making compact identifiers available for long-term use in an ecosystem supporting formal citation of primary scholarly research data. This approach is intended to be robust beyond the operational and funding scope of any one organization, enabling long-term resolution of persistent archived data, whether it is cited in the literature, or is referenced in the web at large. We demonstrate that multiple resolvers with fundamentally different underlying code bases, organizational settings and international alignments, can readily support this approach.

As part of this project we have deployed public, production-quality resolvers using a common registry of prefix-based redirection rules. We believe these products and our approach will be of significant help to publishers, authors and others implementing persistent, machine-resolvable citation of research data in compliance with emerging science policy recommendations and funder requirements.

Introduction

1. Data citability and reuse

Science policy bodies such as CODATA, the Royal Society and the U.S. National Academies have shown significant concern over the past decade about the reliability of published scientific findings and the reusability of research data[1-3]. Policy concerns have been echoed by major funders and by recent statistical and bibliometric research[4-9]. This situation led in 2013-2014 to the development and publication of the Joint Declaration of Data Citation Principles (JDDCP)[10], which has been subsequently endorsed by over 100 scholarly organizations. The JDDCP outlines core principles on purpose, function and attributes of data citations, the first of which is that data should be considered legitimate, citable products of research[11].

The JDDCP require that data become a first-class research object that is archived in persistent stores and is cited just as publications are cited, in all cases where: (1) findings or claims are based on the authors' primary research data; or (2) data from other sources is input to the authors' analysis. JDDCP Principle 3 requires that wherever research findings are based upon data, that data be cited. Principle 4 requires that cited, archived data receive a globally unique, machine-resolvable persistent identifier that appears in the citing article's reference list. This is intended not only to help humans locate data, but to facilitate the development of next-generation mashup tools in an ecosystem based on software agents and searchable research data indexes such as DataMed[12] and OmicsDI[13].

Digital Object Identifiers (DOIs) are already widely used in the publishing world as persistent identifiers for scholarly publications. While DOIs are also used to identify data, they are used to a lesser degree than for articles and even less commonly for biomedical data than for data from other disciplines. Instead, in biomedicine there has been a longstanding practice of employing a prefix plus accession number as a unique identifier. It is not clear that creating DOIs for the billions of existing entities would be worthwhile; even if it did become socially acceptable, financially tractable, and technically achievable, any tangible benefits could be seriously diminished by the complexity and cost of mapping legacy identifiers to the new DOIs.

Our goal in this project was to smooth the path pragmatically for data citation to occur on a wider scale in biomedical research by reducing barriers to its adoption. Consequently, we sought an approach adapted from present practice, which would nonetheless be robust and reliable long-term.

2. How the biomedical research community references data

The life sciences community typically references data entities using their locally-assigned database accession numbers, often contextualizing these local unique identifiers (LUIs) with a prefix corresponding to the assigning authority to clearly indicate the identifier's repository context.

These may be rendered web-resolvable through subsequent incorporation into a durable HTTP URI. The assignment of a *formal*, *registered* namespace prefix avoids identifier collisions, and incorporation into an access URI presents the prefix-identifier combination to a resolver system using web protocols.

In some subdomains such as ontologies, internally-assigned prefixes may be an integral part of the identifier minting process and IDs always appear in their prefixed form[14], for instance, GO:0006915.

We term the result of concatenating a repository-identifying namespace prefix, a colon, and an LUI (<prefix>:<LUI>), a "compact identifier". This approach has been used fairly widely – again, informally – for example to specify International Standard Book Numbers (isbn:<LUI>), Digital Object Identifiers (doi:<LUI>) and other classes of identifier with autonomous assignment.

3. Compact identifier resolvers and meta-resolvers

A core component of persistent identification is redirection, without which it is challenging to provide stable identifiers robust against changes in underlying resource provision. It is thus common practice for bioinformatics data repositories to maintain their own locally-scoped resolvers. A "meta-resolver" is a web

server that can recognize enough about an incoming URL to properly redirect to collection-specific resolvers, based on the assigning authority specified by the incoming URL. Meta-resolvers therefore provide a single host from which to launch URLs containing compact identifiers, by appending the prefix plus LUI to the URL base name of the meta-resolver.

Meta resolvers provide resolution of a common identifier syntax, shield people from the details of and changes to individual resolver access methods, and make it easier to associate compact citations with actionable http links; examples are shown below. Identifiers.org is a type specimen of this class of service, providing several additional services beyond redirection[15]. N2T.net is another such service, with redirection targets based on both individual identifier lookup and prefix-based lookup.

However, until recently, N2T.net could not resolve compact identifiers based on the Identifiers.org prefix set, and Identifiers.org could not resolve the prefix-colon-LUI classical compact identifier format. Nor could either system provide a means to declaratively select from among multiple data providers – for example, from the various hosts of the World Protein Data Bank.

Results

1. Harmonized approach for compact identifier resolution

Our view is that a harmonized resolution approach for compact identifiers must be available to support biomedical data citation, given the sparsity of DOI implementations among biomedical database providers. And to be robust, such an approach requires institutional support at multiple sites in (at least) both Europe and North America.

The specific contribution of our work has been to develop and implement such a robust, harmonized approach to prefix assignment and resolution in the Identifiers.org and N2T.net meta-resolvers, using a common jointly-managed prefix and provider registry. This is meant to be useful to those seeking to implement direct data citation, by providing common methods in both European-based and North-American-based resolver systems, compatible with the widely-used CURIE[16,17] identifier approach; and to serve as a basis for further work in identifier harmonization.

The approach we have developed supports reliable resolution in each case, regardless of the original style of LUIs – whether they are bare numeric (e.g. 9606 in NCBI taxonomy), alphanumeric (e.g. A0A022YWF9 in UniProt), or whether they have a prefix and colon already when issued by the authority (e.g. MGI:80863 in Mouse Genome Informatics). Examples are shown in Table 1 below.

Compact ID	Persistent URL	Provider URL (not persistent)	Resolves to
MGI:80863	http://identifiers.org/MGI:80863	http://www.informatics.jax.org/accession/MGI:80863	JAX Mouse Genome accession number MGI:80863
	http://n2t.net/MGI:80863		
arrayexpress:E-GEOD-2599	http://identifiers.org/arrayexpress:E-GEOD-2599	http://www.ebi.ac.uk/arrayexpress/experiments/ E-GEOD-2599	ArrayExpress accession number E-GEOD-2599
	http://n2t.net/arrayexpress:E-GEOD-2599		
pdb:2gc4	http://identifiers.org/pdb:2gc4	http://www.ebi.ac.uk/pdbe/entry/pdb/2gc4	Protein Data Bank accession number 2gc4
	http://n2t.net/pdb:2gc4		

Table 1. Examples of citable meta-resolver-provided persistent URLs and their corresponding default targets.

2. Optional Provider codes

In many cases, a digital entity can be accessed through multiple different locations. For instance, the NCBI Taxonomy[18], a valuable organism-level classification and nomenclature data resource, is accessible from many sources. As illustrated in Figure 2, if we look at a given taxon such as '9606' (human); it can be accessed directly through the primary source (NCBI), or through thorough copies available at the Ontology Lookup Service (OLS)[19-21] or BioPortal[22,23]. The European Nucleotide Archive also serves NCBI Taxon entities,

primarily to organize and access data about the taxon's corresponding genome assembly. Other examples include the International Nucleotide Sequence Database Collaboration and the Protein Data Bank (PDB)[24]. In many such cases each provider has its own particular virtues in the form of additional services and content.

We have therefore added "provider codes" allowing citation at a specific resolver instance and agreed on a set of formal rules for resolution. The provider code is optionally prepended to the repository namespace and separated with a slash ("/"), thus: (<resolver>/)<provider>/<prefix>:<LUI>. Examples are shown in Table 2 below. Where multiple providers exist, and the provider is not specified, the resolver determines where to resolve the request based on its own rules, e.g., taking availability or other criteria into account.

Compact ID	Provider	Persistent Citable URLs	Resolves to
taxon:9606	If not specified, defaults to NCBI	http://identifiers.org/taxon:9606	NCBI Taxonomy, accession number 9606, NCBI primary
		http://n2t.net/taxon:9606	
taxon:9606	NCBI (**ncbi**)	http://identifiers.org/**ncbi**/taxon:9606	NCBI Taxonomy, accession number 9606, NCBI
		http://n2t.net/**ncbi**/taxon:9606	
taxon:9606	Ontology Lookup Service (**ols**)	http://identifiers.org/**ols**/taxon:9606	NCBI Taxonomy, accession number 9606, OLS
		http://n2t.net/**ols**/taxon:9606	
taxon:9606	NCBO BioPortal (**bptl**)	http://identifiers.org/**bptl**/taxon:9606	NCBI Taxonomy, accession number 9606, BioPortal
		http://n2t.net/**bptl**/taxon:9606	

Table 2. Examples of persistent, citable URLs for a single accession (NCBI Taxon 9606), with default and specified providers.

3. Rules, Registry and Recommendations

Compact Identifiers. A "compact identifier" is a string constructed by concatenating a namespace prefix, a separating colon, and a locally unique identifier (LUI), e.g. pdb:2gc4.

Provider Specification. To specify a specific provider, where multiple providers exist, prepend the provider code and a "/" to the compact identifier, e.g. rcsb/pdb:2gc4.

Provider Default. Where multiple providers exist, and the provider is not specified in the compact identifier, the resolver will determine where to resolve the request based on its own rules, e.g., taking into account uptime availability, regional preference, or other criteria.

Redirect Rule. A URL template associated with the provider code is maintained in the namespace registry, defining how to forward compact identifiers to any specific provider (see 4.2.3 below).

Prefix duplication. Some LUIs (e.g., for Gene Ontology records) may contain embedded namespace prefixes. These are ignored where the result of concatenation would be duplication of the prefix, so a compact identifier example is e.g. GO:0003214, not GO:GO:0003214.

Administration. Prefixes and provider codes can be requested by completing a form at http://identifiers.org/request/prefix. Administrators are currently designated EMBL-EBI and CDL staff.

Resolution. Resolution of compact identifiers is enabled when they are presented as HTTP URIs by prepending the resolution address, e.g. http://identifiers.org/<compactID> or http://n2t.net/<compactID>.

Prefix File. A list of unique namespace prefixes and provider codes in YAML[25,26] format is hosted at https://n2t.net/e/cdl_ebi_prefixes.yaml.

4. Prefix file

The prefix file consists of a sequence of prefix and provider records. The general elements of a "prefix record" follow.

Element: namespace (required)

A string of lowercase letters and digits defining the identifier collection, typically for a given database. The namespace prefix must be unique across the registry.

Element: provider (optional)
A string of lowercase letters and digits defining one provider for an identifier namespace prefix. The provider must be unique across the registry.

Element: alias (optional)

A string of lowercase letters and digits specifying an alternate name for the namespace. The alias must be unique across the registry.

Element: title (required)
A text string containing the full name of the prefix.

Element: homepage (optional)

A URL that leads to a web page with more information about the prefix. If the page contains schema.org tags, the meta-resolver may exploit them for descriptive information.

5. Implementation

We have implemented this approach in both the Identifiers.org resolver (http://identifiers.org), deployed at the European Molecular Biology Laboratory-European Bioinformatics Institute; and the Name-to-Thing resolver (http://n2t.net), deployed at the California Digital Library. These are longstanding public global identifier resolvers with production-quality implementations. They can be used to support machine-resolvable citation of primary research data, in compliance with funder and science policy recommendations. The new features we describe in this article and their joint maintenance have been agreed in a Memorandum of Understanding between CDL and EMBL-EBI. Their support for compact identifiers formalizes longstanding practice in biomedical informatics and supports recent recommendations on biomedical identifiers[27] in a long-term sustainable way.

Conclusions

Compact identifiers are a longstanding informal convention in bioinformatics. In order to be used as globally unique, persistent, web-resolvable identifiers, they require a commonly agreed namespace registry with maintenance rules and clear governance; a set of redirection rules for converting namespace prefixes, provider codes and local identifiers to resolution URLs; and deployed production-quality resolvers with long-term sustainability. An example of formal referencing that leverages a meta-resolver URL is shown with the ArrayExpress reference at the end of this sentence[28].

We have extended prior work of the Identifiers.org team at EMBL-EBI in collaboration with the N2T.net and EZID team at the California Digital Library, and other collaborators, to provide these missing elements.

Taken together, we believe that these implemented measures will facilitate data citation in the scholarly literature, as well as data integration, with an initial focus on the Life Sciences domain. New rules flow into an underlying registry hosted and jointly maintained at the EMBL-EBI, and a public version of the registry is hosted at CDL (https://n2t.net/e/cdl_ebi_prefixes.yaml).

Ongoing support for these services is now agreed in a Memorandum of Understanding between the EMBL-EBI and the California Digital Library. We hope they will be of significant assistance to publishers and others concerned with citation and resolution of biomedical data.

Acknowledgements

This work was funded in part by the European Molecular Biology Laboratory (EMBL); the European Commission within the Research Infrastructures program of Horizon 2020, project numbers 654039 (THOR) and 676559 (Excelerate); and by the U.S National Institutes of Health (NIH) as part of the Big Data to Knowledge (BD2K) program, BioCaddie (U24AI117966), and the Monarch Initiative (R24-OD011883). Work was coordinated by FORCE11 (http://force11.org), a not-for-profit community organization seeking to improve scholarly communication through digital technology.

The authors wish to thank Stephanie Hagstrom of the University of California Library and FORCE11 for her extremely helpful administrative work in connection with the Data Citation Implementation Pilot, in organizing workshops and conference calls, and in coordinating website administration. We also thank Melissa Haendel of Oregon Health and Science University for her helpful comments on the manuscript.

References

1 CODATA/ITSCI Task Force on Data Citation. Out of cite, out of mind: The Current State of Practice, Policy and Technology for Data Citation. Data Science Journal 12, 1-75 (2013) https://doi.org/10.2481/dsj.OSOM13-043.

2 Royal Society. Science as an Open Enterprise. The Royal Society Science Policy Center, London (2012). https://royalsociety.org/topics-policy/projects/science-public-enterprise/report/

3 Uhlir, P. (ed.) Developing Data Attribution and Citation Practices and Standards. National Academies, Washington DC (2012). https://doi.org/10.17226/13564

4 Colquhoun, D. An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science* **1** (2014). https://doi.org/10.1098/rsos.140216

5 Ioannidis, J. A. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* **294**, 218-228 (2005). https://doi.org/10.1001/jama.294.2.218

6 Nissen, S. B., Magidson, T., Gross, K. & Bergstrom, C. T. Publication bias and the canonization of false facts. *eLife* **5**, e21451 (2016). https://doi.org/10.7554/eLife.21451

7 Ramos, M., Melo, J. & Albuquerque, U. Citation behavior in popular scientific papers: what is behind obscure citations? The case of ethnobotany. *Scientometrics* **92**, 711-719 (2012) https://doi.org/10.1007/s11192-012-0662-4

8 Greenberg, S. A. How citation distortions create unfounded authority: analysis of a citation network. *Bmj* **339**, b2680 (2009) https://doi.org/10.1136/bmj.b2680

9 Greenberg, S. A. Understanding belief using citation networks. *Journal of Evaluation in Clinical Practice* **17**, 389-393 (2011) https://doi.org/10.1111/j.1365-2753.2011.01646.x

10 Data Citation Synthesis Group. Joint Declaration of Data Citation Principles. Future of Research Communication and e-Scholarship (FORCE11), San Diego CA (2014). https://doi.org/10.25490/a97f-egyk

11 Altman, M., Borgman, C., Crosas, M. & Martone, M. An introduction to the joint principles for data citation. *Bulletin of the Association for Information Science and Technology* **41**, 43-45 (2015). https://doi.org/10.1002/bult.2015.1720410313

12 Ohno-Machado, L., Sansone, S.A., Alter G., Fore, I., Grethe, J., Xu, H., Gonzalez-Beltran, A., Rocca-Serra, P., Gururaj, A.E., Bell, E., Soysal, E., Zong, N., Kim, H.E. DataMed: Finding useful data across multiple biomedical data repositories. *Nature Genetics*. May 26;49(6):816-819 (2017). https://doi.org/10.1038/ng.3864

13 Perez-Riverol, Y. *et al.* Discovering and linking public omics data sets using the Omics Discovery Index. *Nature Biotechnology*. May 9;35(5):406-409 (2017). https://doi.org/10.1038/nbt.3790

14 Smith, B. *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology* **25**, 1251-1255 (2007). https://doi.org/10.1038/nbt1346

15 Juty, N., Le Novere, N., Hermjakob, H. & Laibe, C. Towards the collaborative curation of the registry underlying Identifiers.org. *Database* : the journal of biological databases and curation 2013, bat017, (2013). https://doi.org/10.1093/database/bat017

16 Birbeck, M. & McCarron, S. CURIE Syntax 1.0, A syntax for expressing Compact URIs: W3C Working Group Note 16 December 2010. (2010). https://www.w3.org/TR/curie

17 Bray, T., Hollander, D., Layman, A., Tobin, R. & Thompson, H. S. Namespaces in XML 1.0 (Third Edition): W3C Recommendation 8 December 2009. (2009). https://www.w3.org/TR/REC-xml-names/

18 Federhen, S. The NCBI Taxonomy database. *Nucleic acids research* **40**, D136-143, (2012). https://doi.org/10.1093/nar/gkr1178

19 Cote, R. *et al.* The Ontology Lookup Service: bigger and better. *Nucleic acids research* **38**, W155-160 (2010). https://doi.org/10.1093/nar/gkq331

20 Cote, R. G., Jones, P., Apweiler, R. & Hermjakob, H. The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC bioinformatics* **7**, 97 (2006). https://doi.org/10.1186/1471-2105-7-97

21 Cote, R. G., Jones, P., Martens, L., Apweiler, R. & Hermjakob, H. The Ontology Lookup Service: more data and better tools for controlled vocabulary queries. *Nucleic acids research* **36**, W372-376 (2008). https://doi.org/10.1093/nar/gkn252

22 Noy, N. F. *et al.* BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research* **37**, W170-173 (2009). https://doi.org/10.1093/nar/gkp440

23 Whetzel, P. L. *et al.* BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic acids research* **39**, W541-545 (2011). https://doi.org/10.1093/nar/gkr469

24 Berman, Helen M., Kleywegt, Gerard J., Nakamura, H. & Markley, John L. The Protein Data Bank at 40: Reflecting on the Past to Prepare for the Future. *Structure* **20**, 391-396 (2012). https://doi.org/10.1016/j.str.2012.01.010

25 Ingerson, B., Evans, C. C. & Ben-Kiki, O. Yet Another Markup Language (YAML) 1.0. (2001). http://yaml.org/spec/history/2001-12-10.html

26 Ben-Kiki, O., Evans, C. & Net, I. d. YAML Ain't Markup Language (YAML™) Version 1.2. (2009). http://www.yaml.org/spec/1.2/spec.html

27 McMurry JA, Juty N, Blomberg N, Burdett T, Conlin T, Conte N, Courtot M, Deck J, Dumontier M, Fellows DK et al. Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. *PLoS Biology* 15(6):e2001414. (2017) https://doi.org/10.1371/journal.pbio.2001414

28 [dataset] Aimone JB, Leasure JL, Perreau VM, Thallmair M. Transcription profiling of rat spinal cord contusion 35 days after injury. ArrayExpress, E-GEOD-2599, 27 March 2012. https://n2t.net/arrayexpress:E-GEOD-2599