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. The former are useful to avoid global identifier collisions when integrating separately managed datasets 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.

In the biomedical domain, the Identifiers.org system supports machine-tractable Web resolution and redirection for names of biomedical digital entities based on a registry of namespaces and a set of resolution and redirection rules. Meanwhile, Name-to-Thing (compactly known as n2t.net) provides resolution of various types of digital entities by managing content directly in its own databases or through prefix-based redirection to the original providers, in a manner akin to Identifers.org.

We report here on significant further work by our team toward making compact identifiers available for long-term use in an ecosystem supporting formal citation of primary research data. This approach is intended to be robust beyond the operational and funding scope of any one organization, enabling long-term resolution of cited persistent data in archives. 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 and rules model. This harmonizes the work of n2t.net, based at the California Digital Library (CDL), University of California Office of the President, and identifiers.org, based at the European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL-EBI). Both resolvers, while derived from independently developed code bases, with different features and objectives, can now uniformly resolve compact identifiers according to our rule set, using a set of common procedures and redirection rules.

We believe these products and our approach will be of significant help to publishers and others implementing persistent, machine-resolvable citation of research data in compliance with emerging science policy body 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 archived in persistent stores and 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 data identifiers, and in many domains outside biomedicine. However, in the case of the over 500 specialized biomedical data archives, they are not commonly used. Instead 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 practically smooth the path for data citation to occur on a wider scale in biomedical research by eliminating barriers to 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 primarily references data entities using locally assigned database accession identifiers, which may be rendered globally unique and machine-actionable through two steps: prefix or namespace assignment and subsequent incorporation into a durable http URI. The assignment of namespace prefix avoids identifier collisions, and incorporation into an access URI presents the prefix-identifier combination to a resolver system using web protocols.

Within a biomedical database, accession identifiers are locally unique identifiers (LUIs). They broadly consist of alphanumeric characters but sometimes include a colon, which delimits an internal prefix assigned locally. Prefixed identifiers are commonplace in some domains and may be an integral part of the identifier 'minting' process[14].

In many cases a specific entity identified through an accession identifier can be accessed through multiple resolving locations, which may for example include formal exact mirroring, consortial data sharing, as well as third party value-added interpretation, transformation or analysis. For instance, the NCBI Taxonomy[15] a valuable organism-level classification and nomenclature data resource is accessible directly through NCBI, or through resources such as the Ontology Lookup Service (OLS)[16-18] or BioPortal[19,20]. Likewise, the Gene Ontology[21] can be retrieved from multiple resources such as AmiGO[21], QuickGO[22], OLS and BioPortal. The Protein Data Bank (PDB) is also available through several providers, each with its own particular virtues in the form of additional services[23]. The approach we have developed is able to accommodate graceful resolution in each case, regardless of the original style of locally unique identifiers--whether they are bare numeric (eg. 9606 in NCBI taxonomy), alphanumeric (eg. 2gc4 in PDB), or whether they have a prefix and colon already when issued by the authority (e.g. MGI:2442292 in Mouse Genome Informatics), etc.

We term the result of concatenating or repository-identifying namespace prefix, with an 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. A formalization of compact identifiers as "CURIEs", a superset of QNAMES, was developed by a W3C working group, although this specification is not a formal standard[24,25].

3. Compact identifier resolvers

We define a "collection resolver" (or just "resolver") to be a web service that accepts URLs with embedded collection identifiers and responds with such things as object access and metadata. Practically every bioinformatics data repository maintains a corresponding collection resolver.

A "meta-resolver" is a web server that can recognize enough about an incoming URL so as to properly redirect to one of potentially several collection resolvers, where the specific resolution is directed through interpretation of prefixes embedded in compact identifiers. A meta-resolver can therefore provide a single host from which to launch URLs containing compact identifiers, which is by appending the prefix plus LUI to the URL base name of the meta-resolver.

A meta-resolver that keeps track of the myriad collection resolver access methods can take an incoming resolution request, lookup the appropriate redirection rule (access method) based on the prefix, and transparently forward the request to the correct collection resolver. This shields people from the details of and changes to collection resolver access methods, and makes it easy to associate compact citations with actionable hypertext links. Identifiers.org is a type specimen of this class of service, providing several additional services beyond redirection [26], with N2T.net being another.

We have modified the syntax of compact identifier presentation to the meta-resolver to parallel the CURIE / QNAMES format, as well as the syntax commonly used in reference sections of biomedical publications for major repositories (e.g. "PMID:26167542" to refer to a 'PubMed' abstract). We have also added "provider codes" allowing citation at a specific resolver instance, and agreed on a set of formal rules for resolution. Lastly we have implemented the rules and registry in both the Identifiers.org and N2T.net resolvers. The registry and rules together promote citability of individual data providers, supported by resolution by a participating meta-resolver. While today this means publishing compact identifiers hosted through either Identifiers.org or N2T.net, the system is nonetheless completely open to adoption by others. The schema employed utilizes an underlying registry hosted and maintained at the EMBL-EBI, which provides additional services and functionality. We propose the use of this prefix registry with an initial focus on the Life Sciences domain.

We have validated this approach through implementation 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 and elsewhere. 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.

We welcome collaborations with those who wish to host mirrors of these resolvers; or to implement the resolution schema and common approach we describe here in their own resolver code.

Rules, Registry and Recommendations

1. **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**.

2. **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**.

3. **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. These criteria should be transparent and publicly available.

4. **Redirect Rule**. A URL-like path 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).

5. **Prefix duplication**. Some LUIs (e.g. for Gene Ontology records) may contain embedded namespace prefixes, e.g. Gene Ontology terms have LUIs beginning with "**GO:**". These are ignored where the result of concatenation would be duplication of the prefix ("**GO:GO:**").

6. **Administration**. Prefixes and provider codes can be requested through normal GitHub "pull requests". Administrators are currently designated EMBL-EBI and CDL staff.

7. **HTTP URI Form**. 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>.

8. **Prefix File**. A list of unique namespace prefixes and provider codes in **YAML** [27,28] format hosted at GitHub: https://github.com/identifiers-org/prefix/blob/master/prefix.yaml.

Compact Identifier resolution

1. Compact Identifiers. A "compact identifier" is a string constructed through the concatenation of a namespace prefix, a separating colon, and a locally unique identifier (LUI) in the collection designated by the namespace prefix, for example, **pmid:16333295** specifies the LUI, **16333295**, in the PubMed collection.

2. Provider Specification. When a collection has more than one provider, prepend a short "provider code" and a "/", for example, epmc/pmid:16333295 specifies the preferred resolver for an individual request to resolve PubMed ID 16333295 as being 'Europe PubMed Central'; rcsb/pdb:2gc4 specifies the preferred resolver for a request to resolve pdb:2gc4 as the Research Collaboratory for Structural Bioinformatics (RCSB)'.

3. Provider Default. 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.

*4. **Prefix file***

4.1 File access

A registry of over 500 curated unique namespace prefixes and provider codes is maintained as a text file in YAML [27,28] format in GitHub, here: https://github.com/identifiers-org/prefix/blob/master/prefix.yaml. Initial contents were loaded from the **identifiers.org** registry (http://identifiers.org/registry), which contains manually curated, high quality data collections [26].

4.2 File format.

The registry is maintained as a text file in GitHub, here:

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

Additions, corrections, and deletions are requested via the standard GitHub issue tracking mechanism.

The registry file consists of a sequence of prefix / provider records for each combination of namespace and provider, as in the following three examples (pmid, epmc/pmid, and flybase).

The general layout of a "prefix record" is:

```
namespace: <required - alphanumeric prefix code for the repository namespace>
provider: <optional - alphanumeric provider code>
redirect: <required - URL forwarding rule>
test: <required - test identifier>
title: <optional - full title of the prefix>
more: <optional - URL for additional information>
note: <optional - text statement>

# comment lines introduced with '#' are allowed
```

4.2.1 Element: scheme

A string of lowercase letters and digits defining the identifier type, typically for a given collection or database. The combination of namespace prefix and provider (below) must be unique across the registry. To indicate a deprecated namespace prefix, end the string with the reserved substring, " - deprecated". This element is required.

4.2.2 Element: provider

A string of lowercase letters and digits defining one provider for an identifier namespace prefix. The combination of namespace prefix and provider (above) must be unique across the registry. If this element is not present, only one provider is assumed. To indicate a deprecated provider, end the string with the reserved substring, " - deprecated".

4.2.3 Element: redirect

A rule specified as a URL-like path defining how to forward compact identifiers to a specific provider. If the literal string "$id" appears in the rule, the redirection target will be formed by replacing the "$id" with the LUI. If no "$id" appears, the LUI will be appended. Unless the rule begins with a fixed URI namespace prefix (eg, "http:" or "https:"), the incoming resolution request's choice of URI namespace prefix will be honored (preserved) in the forwarded request. This element is required.

4.2.5 Element: test

A sample LUI that will be tested before any change to a "redirect" element will be approved (eg, by an automated process). If the result of forwarding the LUI does not ultimately return an HTTP status code in the 200 range, a change in redirect will be denied. This element is required.

4.2.6 Element: title (optional)

A text string containing the full name of the prefix. This element is optional.

4.2.8 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. This element is optional.

4.2.9 Element: note (optional)

A text string containing arbitrary annotations. This element is optional and may be repeated.

Conclusions

Compact identifiers are a longstanding element of practice in biomedical data repositories. In order to be used as globally unique persistent identifiers they require a commonly agreed namespace registry with maintenance rules; 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. 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, ELIXIR, and other collaborators, to provide these missing elements. We hope these tools 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 under the Horizon 2020 THOR and Excelerate programs; and by the U.S.

National Institutes of Health (NIH) as part of the Big Data to Knowledge (BD2K) program. Work was coordinated by FORCE11, 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.

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, doi:http://dx.doi.org/10.2481/dsj.OSOM13-043 (2013).

2 RoyalSociety. Science as an Open Enterprise. (The Royal Society Science Policy Center, London, 2012).

3 Uhlir, P. e. Developing Data Attribution and Citation Practices and Standards. (National Academies, Washington DC, 2012).

4 Colquhoun, D. An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science* **1** (2014).

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

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

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, doi:10.1007/s11192-012-0662-4 (2012).

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

9 Greenberg, S. A. Understanding belief using citation networks. *Journal of Evaluation in Clinical Practice* **17**, 389-393 (2011).

10 Data Citation Synthesis Group. Joint Declaration of Data Citation Principles. (Future of Research Communication and e-Scholarship (FORCE11), San Diego CA, 2014).

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, doi:10.1002/bult.2015.1720410313 (2015).

12 Ohno-Machado, L. *et al.* DataMed: Finding useful data across multiple biomedical data repositories. *bioRxiv* (2016).

13 Perez-Riverol, Y. *et al.* Omics Discovery Index - Discovering and Linking Public Omics Datasets. *bioRxiv* (2016).

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

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

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

17 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, doi:10.1186/1471-2105-7-97 (2006).

18 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, doi:10.1093/nar/gkn252 (2008).

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

20 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, doi:10.1093/nar/gkr469 (2011).

21 Gene Ontology, C. Gene Ontology Consortium: going forward. *Nucleic acids research* **43**, D1049-1056, doi:10.1093/nar/gku1179 (2015).

22 Binns, D. *et al.* QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* **25**, 3045-3046, doi:10.1093/bioinformatics/btp536 (2009).

23 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, doi:10.1016/j.str.2012.01.010 (2012).

24 Birbeck, M. & McCarron, S. CURIE Syntax 1.0, A syntax for expressing Compact URIs: W3C Working Group Note 16 December 2010. (2010).

25 Bray, T., Hollander, D., Layman, A., Tobin, R. & Thompson, H. S. Namespaces in XML 1.0 (Third Edition): W3C Recommendation 8 December 2009. (2009).

26 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, doi:10.1093/database/bat017 (2013).

27 Ben-Kiki, O., Evans, C. & Net, I. d. YAML Ain't Markup Language (YAML™) Version 1.2. (2009).

28 Ingerson, B., Evans, C. C. & Ben-Kiki, O. Yet Another Markup Language (YAML) 1.0. (2001).