10 simple rules for making data web friendly

Nick Juty[1], Julie McMurry[2,3], BioMedBridges Consortium[4], Melissa Haendel[2,3], Carole Goble[5,6], Helen Parkinson[1]

[1]EMBL-EBI, Hinxton, Cambridge, UK. CB10 1SD.
[2]Department of Medical Informatics and Epidemiology and OHSU Library, Oregon Health & Science University, Portland, USA.
[3]Monarch Initiative; http://monarchinitiative.org/
[4]BioMedBridges Consortium; http://www.biomedbridges.eu/partners
[5]School of Computer Science, The University of Manchester, Manchester, UK.
[6]Open PHACTS; http://www.openphactsfoundation.org/

EMBL-EBI

Introduction

Life Science data continues to grow, becoming increasingly available via the Web. Our handling of identifiers for this distributed data, and lack of consideration for how data evolves, has led to phenomena such as 'link rot' (dead links) and 'content drift' (evolved data underlying a stagnant identifier). In addition, poorly conceived strategies for making data available on the Web, built upon poor choices in identifier design and lack of consideration of the global data landscape, is causing downstream integration issues. As a consequence, there is often a need to provide additional and potentially costly procedures for the source-dependent processing of data, or else the provision of alternative mapping solutions. BioMedBridges, in association with an international community of partners, have begun to determine the common technical bridges required to allow data integration across the biological domain. Based on our experience we describe ten simple rules for best practice in the provision and reuse of identifiers for web-based Life Science data [1].

Re-use established identifiers

Never create new identifiers unnecessarily

- *Link using native identifiers*
- *Document relationships to existing entities*
  eg. owl:sameAs, skos:broader

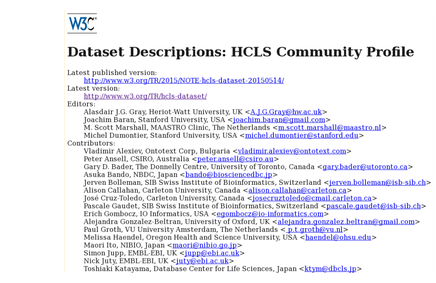
Some desirable identifier characteristics

Unambiguous – *associated with a single entity*
Stable – *identifier to entity relationship*
Versionable – *to allow evolution of entity or metadata*
Persistent – *identifier is never deleted*
Resolvable – *information accessible*
Defined format – *adheres to documented pattern*
Web ready – *contains no reserved characters or special handling*
Documented – *identifier scheme and policy available*

Reference responsibly

Use appropriate referencing and ...

- *Always provide the full URI*
- *Consider a CURIE for narrative online text*
- *Provide additional file to document mappings if necessary*
- *Attempt to ensure that all cross-references remain 'live'*



W3C

Dataset Descriptions: HCLS Community Profile

Latest published version:
http://www.w3.org/TR/2015/NOTE-hcls-dataset-20150514/
Latest version:
http://www.w3.org/TR/hcls-dataset/
Editors:
Alasdair J.G. Gray, Heriot-Watt University, UK <A.J.G.Gray@hw.ac.uk>
Joachim Baran, Stanford University, USA <joachim.baran@gmail.com>
M. Scott Marshall, MAASTRO Clinic, The Netherlands <m.scott.marshall@maastro.nl>
Michel Dumontier, Stanford University, USA <michel.dumontier@stanford.edu>
Contributors:
Vladimir Alexiev, Ontotext Corp, Bulgaria <vladimir.alexiev@ontotext.com>
Peter Ansell, CSIRO, Australia <peter.ansell@csiro.au>
Gary D. Bader, The Donnelly Centre, University of Toronto, Canada <gary.bader@utoronto.ca>
Asuka Bando, NBDC, Japan <bando@biosciencedbc.jp>
Jerven Bolleman, SIB Swiss Institute of Bioinformatics, Switzerland <jerven.bolleman@isb-sib.ch>
Alison Callahan, Carleton University, Canada <alison.callahan@carleton.ca>
José Cruz-Toledo, Carleton University, Canada <josecruztoledo@cmail.carleton.ca>
Pascale Gaudet, SIB Swiss Institute of Bioinformatics, Switzerland <pascale.gaudet@isb-sib.ch>
Erich Gombocz, IO Informatics, USA <egombocz@io-informatics.com>
Alejandra Gonzalez-Beltran, University of Oxford, UK <alejandra.gonzalez.beltran@gmail.com>
Paul Groth, VU University Amsterdam, The Netherlands <p.t.groth@vu.nl>
Melissa Haendel, Oregon Health and Science University, USA <haendel@ohsu.edu>
Maori Ito, NIBIO, Japan <maori@nibio.go.jp>
Simon Jupp, EMBL-EBI, UK <jupp@ebi.ac.uk>
Nick Juty, EMBL-EBI, UK <juty@ebi.ac.uk>
Toshiaki Katayama, Database Center for Life Sciences, Japan <ktym@dbcls.jp>

Document your practices

Document the identifiers you create, and those you use from other providers. Include HCLS Dataset Description [2], e.g.

- *Entity scope*
- *licence*
- *Namespace and mapping*
- *Machine-readable formats*

Follow identifier conventions

Follow conventions when minting new identifiers

- *No internal whitespace*
- *Avoid 'date' or 'exponent' misinterpretation*
- *Specify a 'pattern' for identifiers*

Name	Namespace	Definition
UniProt Knowledgebase	uniprot	The UniProt Knowledgebase (UniProtKB) is a comprehensive resource for protein sequence and functional information with extensive cross-references to more than 120 external databases. Besides amino acid sequence and a description, it also provides taxonomic data and citation information.
BiGG Metabolite	bigg.metabolite	BiGG is a knowledgebase of Biochemically, Genetically and Genomically structured genome-scale metabolic network reconstructions. It more published genome-scale metabolic networks into a single database with a set of stardized identifiers called BiGG IDs. Genes in the BiGG models are mapped to NCBI genome annotations, and metabolites are linked to many external databases (KEGG, PubChem, and many more). This collection references individual metabolotes.
UMLS	umls	The Unified Medical Language System is a repository of biomedical vocabularies. Vocabularies integrated in the UMLS Metathesaurus include the NCBI taxonomy, Gene Ontology, the Medical Subject Headings (MeSH), OMIM and the Digital Anatomist Symbolic Knowledge Base. UMLS concepts are not only inter-related, but may also be linked to external resources such as GenBank.
InChIKey	inchikey	The IUPAC International Chemical Identifier (InChI, see MIR:00000383) is an identifier for chemical substances, and is derived solely from a structural representation of that substance. Since these can be quite unwieldly, particularly for web use, the InChIKey was developed. These are of a fixed length (25 character) and were created as a condensed, more web friendly, digital representation of the InChI.
SEED Subsystem	seed	This cooperative effort, which includes Fellowship for Interpretation of Genomes (FIG), Argonne National Laboratory, and the University of Chicago, focuses on the development of the comparative genomics environment called the SEED. It is a framework to support comparative analysis and annotation of genomes, and the development of curated genomic data (annotation). Curation is performed at the level of subsystems by an expert annotator, across many genomes, and not on a gene by gene basis. This collection references subsystems.
SEED Compound	seed.compound	This cooperative effort, which includes Fellowship for Interpretation of Genomes (FIG), Argonne National Laboratory, and the University of Chicago, focuses on the development of the comparative genomics environment called the SEED. It is a framework to support comparative analysis and annotation of genomes, and the development of curated genomic data (annotation). Curation is performed at the level of subsystems by an expert annotator, across many genomes, and not on a gene by gene basis. This collection references subsystems.
Taxonomy	taxonomy	The taxonomy contains the relationships between all living forms for which nucleic acid or protein sequence have been determined.
BiGG Model	bigg.model	BiGG is a knowledgebase of Biochemically, Genetically and Genomically structured genome-scale metabolic network reconstructions. It more published genome-scale metabolic networks into a single database with a set of stardized identifiers called BiGG IDs. Genes in the BiGG models are mapped to NCBI genome annotations, and metabolites are linked to many external databases (KEGG, PubChem, and many more). This collection references individual models.

http://identifiers.org/registry

Assign namespaces

Provide a namespace and mapping to allow local identifiers to be used globally

- *Document the namespace used*
- *Document the mapping to a resolving location*
- *Register namespace where appropriate*

	A) Compact URI (CURIE) <prefix>:<local resource identifier>		B) URIs <namespace>:<local resource identifier>		C) Access URLs
In-house	1) Native	ZFIN:**ZDB-GENE-980526-166**	http://zfin.org/**ZDB-GENE-980526-166**	no redirect	http://zfin.org/**ZDB-GENE-980526-166**
In-house	2) PURL	UniProt:**A0A022YWF9**	http://purl.uniprot.org/uniprot/**A0A022YWF9**	→	http://www.uniprot.org/uniprot/**A0A022YWF9**
In-house	3) Mirroring	ENSEMBL:**ENSMUSG00000033577**	http://www.ensembl.org/id/**ENSMUSG00000033577**	↗	http://uswest.ensembl.org/Mus_musculus/Gene/Summary?g=**ENSMUSG00000033577**;r=9:80165031-80311729;redirect=no http://asia.ensembl.org/Mus_musculus/Gene/Summary?g=**ENSMUSG00000033577**;r=9:80165031-80311729;redirect=no
3rd party	4) identifiers.org	BioSample:**SAMN03491154**	http://identifiers.org/biosample/**SAMN03491154**	→	http://www.ncbi.nlm.nih.gov/biosample/**SAMN03491154**
3rd party	5) DOI	doi:**10.5281/zenodo.18003**	http://dx.doi.org/**10.5281/zenodo.18003**	→	https://zenodo.org/record/18003

Provide URIs

Implement intuitive and resolvable URIs

- *Avoid format, file extension and administrative parameters*
- *Provide a landing page*
- *Consider CURIEs for display*

CURIE

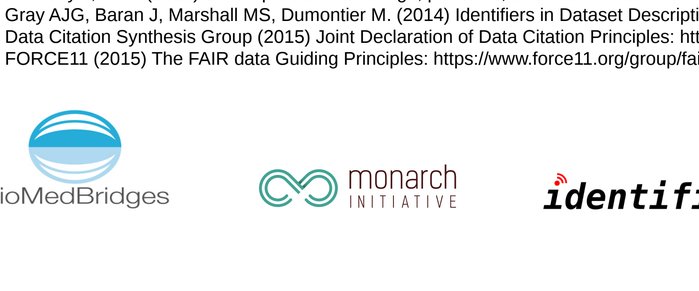
Compact URI = <**Prefix**>:<**Local Identifier**>
Prefix is defined and documented
Allows deterministic expansion to resolving location
Mnemonic useful in human communication
Suitable for convenient display to users

ZFIN:ZDB-GENE-980526-166

Avoid embedded semantics

Would embedding semantic information in an identifier make it prone to obsoletion in an evolving data domain?

- *Define the entity being identified*
- *Provide metadata about the entity*
- *Semantic information acceptable in some cases (InChI)*



UniProtKB - P29358 (143B_BOVIN)

Basket

This entry is obsolete

On October 25, 2004 this entry became obsolete. It can now be found as secondary accession in P68250 and P68251. [List]
For previous versions of this entry, please look at its history.

UniProt

History

Previous versions for UniProtKB entry: P12345

Download

	Versions		Info	Releases
	Entry	Sequence	Entry name	Database
	100 .txt	2 .fasta	AATM_RABIT	Swiss-Prot
	99 .txt	2 .fasta	AATM_RABIT	Swiss-Prot
	98 .txt	2 .fasta	AATM_RABIT	Swiss-Prot
	97 .txt	2 .fasta	AATM_RABIT	Swiss-Prot

For a permanent link to this page, which will not change with the next release of Ensembl, use:

http://Sep2015.archive.ensembl.org/Mus_musculus/Gene/Summary?g=ENSMUSG00000033577;r=9:80165031-80311729

We aim to maintain all archives for at least two years; some key releases may be maintained for longer

project Ensembl

Never reassign

Never re-assign an identifier to a different record. (cf. Ghost Busters 'don't cross the streams')

- *Provide a list of all obsolete identifiers*
- *Create new identifier on record merge*
- *Create new identifiers for demerged*
- *Always provide links for the above*

Use versioning

Document change history at the record level, and/or provide versioned identifiers

- *Tombstone page for obsolete*
- *Use '.' to version local identifier*
- *'provide links to latest record*

Make URIs clear

Allow users to easily identify the URI for reuse

- *Advertise a permanent link*
- *Clearly label out-dated records*
- *Consider providing 'cite this' button*

Summary

A plethora of issues undermine the identification of data in the Life Sciences, hampering our ability to integrate the ever increasing amounts of data being produced. We have outlined best practices for data providers, as well as guidance for data integrators and redistributors, to begin the process of solving at least some of these issues. We hope that the information provided here will also be useful to data generators and end users, allowing them to appreciate the additional complexity of making data available in a web environment. We anticipate that, over time, improved tooling will become available, lowering the barriers to adoption. We also wish to acknowledge the work of several groups internationally[3.4], who are also converging on standards that will be applicable to identifiers.

References

1. McMurry J, *et al.* (2015) 10 Simple rules for design, provision, and reuse of identifiers for web-based life science data: Zenodo. 10.5281/zenodo.31765
2. Gray AJG, Baran J, Marshall MS, Dumontier M. (2014) Identifiers in Dataset Descriptions: HCLS Community Profile. In: HCLS Community Profile: http://www.w3.org/2001/sw/hcls/notes/hcls-dataset/
3. Data Citation Synthesis Group (2015) Joint Declaration of Data Citation Principles: https://www.force11.org/datacitation
4. FORCE11 (2015) The FAIR data Guiding Principles: https://www.force11.org/group/fairgroup/fairprinciples

BioMedBridges · monarch INITIATIVE · identifiers.org · Open PHACTS · elixir · EMBL

EMBL-EBI
Wellcome Trust Genome Campus
Hinxton, Cambridgeshire, CB10 1SD, UK

Tel. +44 (0) 1223 494 444
juty@ebi.ac.uk
www.ebi.ac.uk