Introduction

Life science data is evolving to be ever larger, more distributed, and more natively web-based. However, our collective handling of identifiers has lagged behind these advances. Diverse identifier issues (for instance “link rot” and “content drift” [1]) have hampered our ability to integrate data and derive new knowledge from it. Optimizing web-based identifiers is harder than it appears and no single scheme is perfect: Identifiers are reused in different ways for different reasons, by different consumers. Moreover, digital entities (e.g., files), physical entities (e.g., biosamples), and descriptive entities (e.g., ‘mitosis’) have different requirements for identifiers. Nevertheless, there is substantial room for improvement throughout the life sciences and several other groups have been converging on identifier standards that are broadly applicable [2, 3, 4, 5].

Building on these efforts and drawing on our experience, we focus on the use case of large-scale data integration: we outline the identifier qualities and best practices that we feel are most important in this context. Specifically, we propose actions that providers of online databases (repositories, registries, and knowledgebases) should take when designing new identifiers or maintaining existing ones (**Rules 1-9, Box 1**). In **Rule 10**, we conclude with guidance to data integrators and redistributors on how best to reference identifiers from these diverse sources. This article may also be useful to data generators and end users as it offers insight into the issues associated with data provision in a web environment. We call upon data providers to take a long-term view of their entities’ scope and lifecycle, and to consider existing identifier platforms and services [6].

Throughout this document, the keywords "must", “should”, "recommended", etc are used here interpreted as described by the W3C[7]. Terms that appear in fixed-width font are also defined in the supplemental glossary (**Table S2**).

Rule 1. Use established identifiers
Rule 2. Design identifiers for use by others
Rule 3. Help local identifiers travel well: document Prefix and Namespace
Rule 4. Opt for simple durable web resolution
Rule 5. Avoid embedding meaning
Rule 6. Make URIs clear and findable
Rule 7. Implement a version management policy
Rule 8. Do not re-assign or delete identifiers
Rule 9. Document the identifiers you issue and use
Rule 10. Reference responsibly

Box 1. A summary of the 10 simple rules

Rule 1: Use established identifiers

If you manage an online database (repository, registry, or knowledgebase), you are likely to have new entities to identify. You might also issue alternate identifiers for external entities, for example to reduce risks posed by dependency on an outside source or to identify meaningful differences in an entity, its state, or its representation.

If you must create your own *alternate identifiers*, you must document the relationship between the existing and alternate identifiers using established properties such as ro:derives_from, owl:sameAs, or skos:broader. If the motivation to create a new identifier is based upon the need to provide factual corrections of content, it is best to work with the database-of-origin to fix the source record rather than create a new one. Wherever the 1:1 relationship of identifier:entity breaks down, costly mapping problems are created. Wherever possible, reference well-established identifiers (even problematic ones; see Rule 10) rather than creating new ones.

Rule 2: Design identifiers for diverse uses by others

Pre-existing identifiers should be referenced without modifications (see **Rule 10**). However, when new local identifiers are necessary, there are some design decisions that can facilitate their use in diverse contexts (spreadsheets, other databases, web applications, publications, etc.).

We use the term `Local Resource Identifier`[8] (`LRI`, **Box 2**) to mean a publicly available identifier that is unique within a single dataset. `Local Resource Identifiers`:

- Must comprise only printable ASCII characters without whitespace. This guards against corruption and mistranscription in many contexts.
- Should contain both letters and numbers. This avoids misinterpretation as numeric data (e.g. truncation of leading zeros in spreadsheets).
- Should avoid problem patterns; this avoids misinterpretation whether as dates, exponents in spreadsheets [9], or unintended words.
- Should adhere to a fixed, documented case convention, preferably one that is case insensitive; this avoids accidental collisions.
- Must adhere to a formal pattern (regular expression); this facilitates but does not guarantee validation and retrieval from scientific text. Consider a fixed length of 8-16 characters (according to the anticipated number of required LRIs). A pattern may be extended if all available identifiers are issued, but existing identifiers must not be changed. To minimize global LRI collisions, it is considerate to tightly specify your pattern (e.g. using two or more fixed letters at the start).
- Should ideally not contain '.' except to denote version where appropriate (see **Rule 7**)

Two small considerations also make LRIs well suited for others to use in user-friendly compact notation and semantic web. We therefore recommend that LRIs:

- Should not contain ':', a reserved character for CURIE parsing **(Box 2)**
- If additional delimiters (other than ':' and '.') are needed, prefer '-'. This guards against certain `CURIE` parsers splitting inappropriately.

Rule 3: Help local identifiers travel well by documenting Prefix and Namespace

Data does not live in silos: it is reused, broken into parts and integrated with other data, most notably in database external references (aka "XRefs"), in the Semantic Web, and in publications (articles and research datasets). The Local Resource Identifier (**Box 2**) alone is insufficient for these tasks because it is only guaranteed to be locally unique. For instance, the LRI "9606" corresponds to numerous entities whose local accessions are based on simple digits, including: a Pubmed article, a CGNC gene, a PubChem chemical, as well as an NCBI taxon, a BOLD taxon, and a GRIN taxon.

Despite its vulnerabilities, the location-based identifier scheme (`URI,` Box 2) is the best available identifier form for machine-driven global data integration because it is a) widely adopted and b) its uniqueness is ensured by a single well-established name-granting process (`DNS`). Juty et al. [10] summarise why name-based global identifier schemes (e.g. `URN`s) have had poor uptake by comparison.

The length of `URIs` (**Box 2**) can make them unwieldy for tasks involving human readability, even within structured machine-parsable documents[11]. `Compact URIs (CURIEs`[12], Box 1) are a well established convention in such contexts (e.g. `JSON-LD` and `RDFa`) as they enable `URIs` to be understood and conveniently accessed. `CURIEs` complement URIs, rather than replace them. Therefore document the prefix which others may use to abbreviate your identifiers for human readability, wherever needed. If you are a database provider, it is in your best interests to document a) the `prefix` (**Box 2**) that you would like others to use and b) its binding to a `resolving namespace` (**Box 2**). Your chosen prefix should be unique, at least among datasets that are likely to be used in the same context. To facilitate this, we strongly recommend that you register your `prefix` and `resolving namespace;` **Table S3** contains a list of registries that may be suitable depending on the kind of data.

Box 2. Local and Global Identifier Terminology

An **identifier** is a sequence of characters that identifies an entity.

- **Local Resource Identifier (LRI)** is an identifier that is unique within the scope of a single database.
  - Databases and library systems often refer to the LRI as an 'Accession Number'.
  - LRI formats vary by provider and may have subparts; however subparts are non-uniform and therefore not described here. For example, a LRI may be opaque (e.g. A0A022YWF9) or recognizable (e.g. ZDB-GENE-980526-388)
- **Global identifier** is an **identifier** that is guaranteed to be globally unique
  - **Uniform Resource Identifier (URI)** is an identifier that is uniform; it is an ASCII string that uniquely identifies a resource. In this paper, by URI we mean only those URIs of type HTTP, HTTPS, FTP, etc. that actually resolve to (provide or redirect to) a webpage containing information about the identified entity. An example of a URI is http://zfin.org/ZDB-GENE-980526-388.
  - When referring to **compact URIs (CURIES)**, we mean an identifier comprised of **<Prefix>:<LRI>** wherein **prefix** is deterministically expandable to **a resolving namespace** (see below) which *alone* is the basis for the CURIE's global uniqueness. An example of a CURIE is ZFIN:ZDB-GENE-980526-388
- A **resolving namespace** is a sequence of characters which, when prepended to the **LRI**, yields the **URI.** In the ZFIN example above, the **prefix** is ZFIN and the resolving namespace is http://zfin.org/

See also **Fig 1** and glossary (**Table S2**) for additional terms and concepts.

Table 1. Desirable characteristics for database identifiers in the life sciences

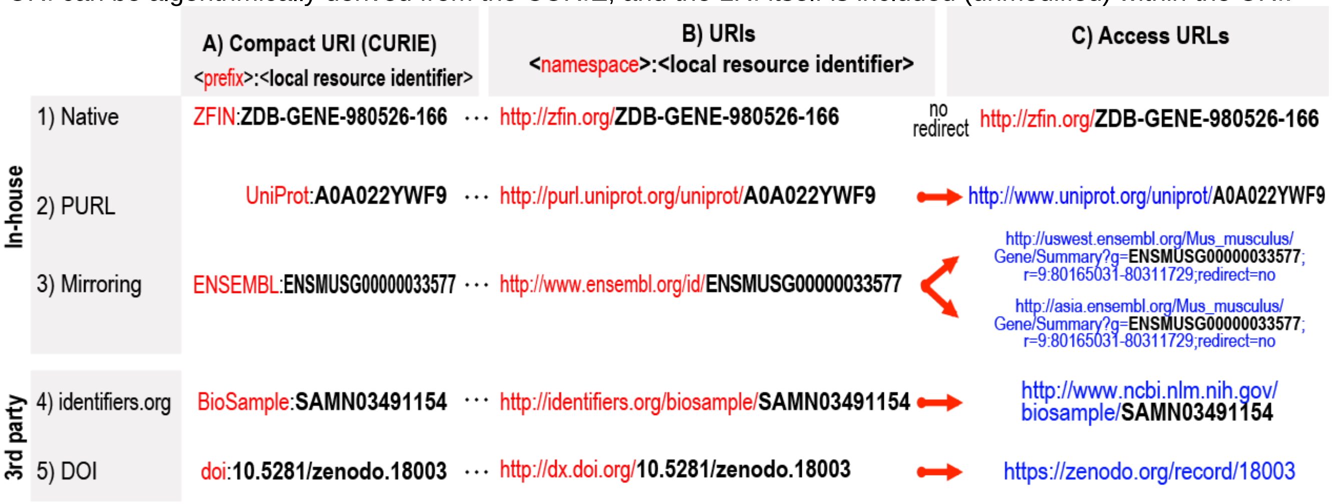
Characteristics	Definition	Rationale/impact on data integration
Unambiguous	One LRI MUST be associated to no more than one entity *locally.* One URI MUST be associated to no more than one entity *globally*	Avoids collisions that result in integrating on the wrong entity
Unique	One entity SHOULD be identified by no more than one LRI and no more than one URI	Eliminates costly mapping problems and avoids false negatives if identifier equivalence cannot be determined
Stable (identifier)	The identifier MUST stay the same over time[a]	Avoids link rot
Stable (entity)	Identifier MUST NOT be reassigned to an altogether different entity, though the entity may evolve provided a change history is documented	Avoids integrating on the wrong entity
Version-documented	If the entity's definition or essential metadata changes, (Rule 7) the identifier MUST be versioned and/or change history documented	Avoids integrating on the wrong entity state
Persistent	The identifier MUST NOT be deleted	Avoids link rot
Web-resolvable	The URI MUST be resolvable to a web address where the data or information about the entry can be accessed	Avoids the unnecessary proliferation of resolvable identifiers issued by third parties (for entities that are not resolvable or identified in their native context)
Convertible	The LRI and its URI counterpart MUST be inter-convertible by prepending resolving namespace	Avoids the need for special handling of edge cases when integrating data at scale
Defined	The identifier MUST each adhere to a formal pattern (e.g. regular expression)	Facilitates validation and extraction from text
Web-friendly	The LRI MUST be of a format that does not need special handling when used in URLs and common exchange formats (e.g. XML)	Avoids potential points of failure due to malformed URL, XML, etc.
Free to assign	The identifier SHOULD be assigned at no cost to individuals depositing data in a repository	Encourages data providers to deposit data
Open access and use	The identifier SHOULD be able to be transparently referenced and actioned (e.g. in a public index or search) anywhere by anyone and for any reason. Restrictions on associated data may apply but are not recommended.	Enables integration on the basis of scientific merit, rather than on the restrictions of the license
Documented	The identifier scheme MUST be documented	Encourages consistent use of existing identifiers. Decreases identifier proliferation.

[a] Berners-Lee T. Cool URIs don't change. 1998. [Cited 2015 May 15]. [Internet]. Available: http://www.w3.org/Provider/Style/URI

Rule 4: Opt for simple, durable web resolution

If you are a database provider, you must implement a `resolving namespace` (**Fig. 1 panel B**) for local identifiers to be "resolvable" to a web page. Use best practices to implement `content negotiation` for different encodings of your data [2], and provide direct access to data, metadata, and persistence statements [13]. If you choose to outsource to a resolver service, use an approach that is JDDCP approved [2] (e.g. DataCite, CrossRef, Identifiers.org, Handle.net, PURL, EPIC, ARK) and be mindful of your constraints regarding cost, metadata ownership, turnaround time, etc. (See **Text S5** for a more comprehensive list of considerations.) If you have the resources to support your own persistent URIs, design these to be simple: Omit anything that is likely to change or lapse, including administrative details (e.g. grant name) or implementation details such as file extensions ('`resource.html`'), query strings ('`param=value`'), and technology choices ('`.php`'). The compact URI approach can work with any resolver(s): see for instance examples 4 and 5 in **Figure 1**. By choosing a single `namespace` per database, you make it possible for others to resolve your identifiers simply (**Fig. 1 panel A**). If multiple resolvers are used, each must have a corresponding prefix (e.g. KEGG-path: http://www.kegg.jp/dbget-bin/www_bget?path: vs KEGG-ko: http://www.kegg.jp/dbget-bin/www_bget?ko:) Occasionally, the **`resolving namespace`** is the same as the homepage (e.g. http://zfin.org/ in **Fig. 1**). In all cases, the **`resolving namespace`** must be exactly as it appears in the URI: it must include the protocol (e.g. http://) and, if applicable, trailing slash or other delimiters.

Fig. 1. Examples of provisioning resolvable URIs:
Compact URIs (CURIEs) (Panel A), URIs (Panel B) and Access URLs (Panel C) with in house examples from (ZFIN, UniProt, and ENSEMBL and 3rd party resolver examples using identifiers.org and DOI. In each case, the URI can be algorithmically derived from the CURIE, and the LRI itself is included (unmodified) within the URI.



		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
	2) PURL	UniProt:A0A022YWF9	···	http://purl.uniprot.org/uniprot/A0A022YWF9	→	http://www.uniprot.org/uniprot/A0A022YWF9
	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
	5) DOI	doi:10.5281/zenodo.18003	···	http://dx.doi.org/10.5281/zenodo.18003	→	https://zenodo.org/record/18003

Rule 5: Avoid embedding meaning

The structure and scope of collections evolve, as does scientific understanding; minimizing the meaning embedded in identifiers makes them less vulnerable to obsoletion. In human genetics many genes were initially identified based on disease association. Later the identification, nomenclature, and function of genes were separated into different activities. It is still possible to embed precision within an identifier: for instance, an InChI string both identifies a chemical entity and defines its structure. Hence, meaning can be embedded where it is indisputable, unchangeable and/or useful to the data consumer (e.g. computer-processable). These rules of thumb apply especially to LRIs but also to the path of the URIs (see **Rule 4**).

When assigning identifiers, define what kind of entity is being identified. This information must be provided as an available description and encoded where possible, using metadata landing pages [2][13].

Rule 6: Make URIs clear and findable

Make URIs obvious to users, especially where these differ from access URLs or application pages. For instance, at the record-level, advertise the "permanent link" together with a statement about persistence. E.g. "The permanent link to this page, which will not change with the next release of Ensembl is: http://Jul2015.archive.ensembl.org/Mus_musculus/Gene/Summary?g=ENSMUSG00000033577;r=9:80165031-80311729;redirect=no We aim to maintain all archives for at least two years; some key releases may be maintained for longer"

For archived records that are *out of date,* make this clear to the user and provide a link to the updated version (see http://www.uniprot.org/uniprot/P12345.1, for instance). Although it is good practice for each database website to include general citation guidance for users, it is ideal to provide a "cite this" button at the level of each record.

Rule 7: Implement a version-management policy

Changes in data resources impact how they can be referenced and used. If you issue identifiers, document the change history for the resource (see also **Rule 8**), or version the identifier itself, or do both and document these.

Explicit versioning is recommended if prevailing use of an unversioned identifier results in "breaking changes" (e.g., a change in the hypothesized cause of a disease). However, if new information about the entity emerges slowly and the changes are "non-breaking", it is reasonable to instead maintain a machine-actionable change history wherein the changes are also categorized. Versioning and change history work well together, especially when multiple types of changes overlap. Even when previous records are removed, the URI should continue to resolve, but to a "tombstone" page (**Rule 8**).

A summary of versioning recommendations follows in **Tables 2a and 2b** below. See Kratz et al. [15] for a more in-depth discussion of change management considerations. If you version identifiers at the level of the individual record, you must version in the LRI after the dot per UniProt in **Table 2a**; this provides continuity in your site and also enables a single CURIE prefix to be used with any version: UniProt:P12345.3 → http://www.uniprot.org/uniprot/P12345**.3.**

Table 2a. Recommendation for record-level versioning with URIs

Recommendation	**Example** (for clarity, LRI only is shown)
Version information should follow after a dot	P12345**.3**
Base resource must resolve (302 redirect) to most recent version	P12345
Base resource should be deterministically convertible from version	P12345**.1** to P12345
Older versions must resolve	P12345**.1**
Illegal or invalid version must produce an informative error message	P12345**.302**
Link from older version to current version must be provided	P12345**.3**
A list of all previous versions should be available	P12345 (see 'history' tab in user interface)
Two versions (or dates) should be comparable	http://www.uniprot.org/uniprot/P12345?version=*

Table 2b. Recommendation for database-release versioning with URIs

Recommendation	Versioning may be done in the namespace (and ideally prefix)
URI example	**http://Jul2015.archive.ensembl.org/Mus_musculus/Gene/Summary?g=**ENSMUSG00000033577
CURIE example	**ENSEMBL-2015-07-MUSG:**ENSMUSG00000033577

Rule 8: Do not reassign or delete identifiers

Identifiers generated and publicly advertised must never be reassigned to a different record or deleted. If you issue identifiers, consider their full lifecycle: there is a fundamental difference between identifiers which point to experimental datasets (GenBank/ENA/DDBJ, PRIDE, etc.) and identifiers which point to a current understanding of a biological concept (Ensembl Gene, UniProt record, etc.). While experimental records are less likely to change, concept descriptions may evolve rapidly; even the nature and number of the relevant metadata fields change over time. Moreover, the very notion of identity is often strongly impacted by relationships (e.g., between concepts or processes).

Extensive changes cannot be captured with numerical suffixing alone. For instance, taxonomists may split or merge species, pathologists may split or merge diseases, or hypothesized entities may be proven not to exist (e.g. vaccine-induced autism). Global initiatives (**Text S1**) are actively exploring identifier strategies for such use cases. In the meantime, consider **Table 3** recommendations.

Table 3. Recommendations for identifier lifecycle management

Recommended handling	**Example**
Obsoletion: If an entry has been removed or deprecated, the original identifier must still resolve to a `‘tombstone page’`. Reasons for obsolescence should be indicated. If the obsoleted ID is replaced by another ID, the replacement must be present and also described as automatic or suggested, preferably using the ontology properties `iao:replaced_by` and `obo:consider`, respectively. The obsoleted ID must never be reassigned to another entity. A list of obsoleted IDs should be maintained.	Single obsoleted identifier: http://www.uniprot.org/uniprot/A0AV18 List of obsoleted identifiers: uniprot.org/help/deleted_accessions
Merging: When two or more identifiers are merged, a new recipient identifier should be designated as the primary (citable) one and should contain information about the legacy identifiers it encompasses. Any legacy identifiers should continue to resolve via redirection to the primary identifier.	UniProt entries Q57339 and O08022 have been merged into Q00626. Q57339 and O08022 are redirected to Q00626.
Splitting: If an identifier is split (demerged) into two or more new ones, new identifiers should be assigned to all the new entries. The legacy identifier must be obsoleted, must resolve, and should provide a warning and pointers to the new ones as per above.	UniProt entry P29358 has been split into P68250 and P68251. P29358 displays a warning and links to the demerged entries: http://www.uniprot.org/uniprot/P29358

Rule 9: Document the identifiers you issue and use

The global-scale identification cycle is a shared responsibility and provider/consumer roles often overlap in the context of data integration. Whether you issue your own identifiers or just reference those of others, you must document your identifier policies. **Supplemental Table S6** provides a set of questions that data providers and re-distributors can use to develop such documentation. Documentation should be published alongside and/or included together in a dataset description, as outlined in the recommendations for Dataset Descriptions developed by the W3C Semantic Web in the Health Care and Life Sciences Interest Group [16]. For examples of such documentation see ChEMBL[17] and Monarch[18] ; the format may vary.

Rule 10: Reference responsibly

The final rule describes referencing recommendations for data redistributors: data aggregators, who collect information from different sources and re-display it; data publishers, who disseminate scientific knowledge through publications; and online reference material.

When database identifiers are referenced in narrative online text, they should always be hyperlinked to their URIs or to metadata containing their URIs [e.g. 19, 20]. Where cross-references are displayed to humans (e.g. as in Monarch disease overview pages) consider using a CURIE notation whose `prefix` (see **Rule 2**) is as assigned by the data provider. Where machine parsing is intended (e.g. Monarch metadata landing pages), any CURIEs must be given together with a machine-parsable definition of the prefix-to-source mapping e.g. Monarch CURIE map. `Access URLs` are volatile (see Rule 4) and must not be used for referencing. This is especially relevant for bundled and transitive references to identifiers, such as those found in published bioinformatics pipelines (for example, Research Object or Galaxy).

Where there exists no provider-issued or commonly-used prefix for a dataset, the data distributor should select one, preferably after discussion with the data provider. Similar documentation best practices apply to data providers and redistributors (**Table S6**).

Redistributors of data should monitor their references to other sources; any 'dead' links should be reported to the original data provider. If the original provider does not fix the broken link, the reference to it should be marked obsolete both visibly (for user interaction/interpretation), and within any accompanying metadata (for computational interaction/propagation).

Conclusion

Better identifier design, provisioning, documentation, and referencing can address many of the identifier problems encountered in the life science data cycle. We recognize that improvements to the quality, diversity, and uptake of identifier tooling would lower barriers to adoption of these rules. We will undertake to address these gaps in the relevant initiatives (**Text S1**). We also recognize the need for formal software-engineering specifications of identifier formats and/or alignment between existing specifications and hope that this paper can catalyze such efforts.

Acknowledgments

J McMurry, T Burdett, N Juty, and S Jupp were supported in part by the BioMedBridges project, which is funded by the European Union Seventh Framework Programme within Research Infrastructures of the FP7 Capacities Specific Programme, grant agreement number 284209. EMBL-EBI core funds supported H Parkinson, MJ Martin, J McEntyre, H Hermjakob. J Malone. ELIXIR core funds supported N Blomberg, R Jimenez. The European Commission provided additional support for Simon Jupp under grant number 601043 ("DIACHRON") and for N Juty and H Hermjakob under grant number 312455 ("Infrastructure for Systems Biology - Europe (ISBE)"). The Drug Disease Model Resources grant number DDMoRe 115156 ("Innovative Medicines Initiative") supported C. Laibe. Support was also received from the following BBSRC grants: BB/L005050/1 ("ELIXIR-UK, Manchester") for SA Sansone, A Gonzalez-Beltran and C Goble; BB/M013189/1 ("DMMCore") for C Goble, J Snoep, and N Stanford; BB/K019783/1 ("Continued development of ChEBI") and BB/M006891/1 ("EMPATHY") for N Swainson; BB/M017702/1 ("SYNBIOCHEM") for N Swainson and D Fellows; BBS/E/B/000C0419 ("A systems approach to understanding lipid, Ca2+ and MAPK signalling networks") for N Le Novère; BB/L005069/1 ("ELIXIR-UK, Oxford") for SA Sansone, A Gonzalez-Beltran and P Rocca-Serra. NIH support was provided from the following grants: U41HG007822 ("UniProt") for MJ Martin; U24AI117966-01 ("bioCADDIE") for SA Sansone, A Gonzalez-Beltran and P Rocca-Serra; U54AI117925 ("CEDAR") for SA Sansone, A Gonzalez-Beltran and P Rocca-Serra; R24OD011883 ("Semantic LAMHDI") for CJ Mungall, MA Haendel, JA McMurry and NL Washington; NHGRI P41HG002273-09 ("Gene Ontology Consortium") for CJ Mungall. Additional support for CJ Mungall and NL Washington was received from the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under [Contract No. DE-AC02-05CH11231].

The authors wish to thank Mary Todd Bergman, Ewan Birney, Fiona Cunningham, Richard Cyganiak, Adam Faulconbridge, Andrew M Jenkinson, Sirarat Sarntivijai, Stephanie Suhr, Nicole Vasilevsky, and Eleanor Williams for their valuable feedback and suggestions. We also wish to thank the BioMedBridges Scientific Advisory Board for the suggestion to address this important issue and the reviewers for their constructive comments.

References

1. Van de Sompel H, Sanderson R, Shankar H, Klein M (2014) Persistent Identifiers for Scholarly Assets and the Web: The Need for an Unambiguous Mapping. International Journal of Digital Curation 9: 331-342. doi: 10.2218/ijdc.v9i1.320
2. Data Citation Synthesis Group (2014) Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11. [Internet] Available: [https://www.force11.org/datacitation. Accessed: 23 September 2015.
3. Altman M, Crosas M (2013) The Evolution of Data Citation: From Principles to Implementation. IASSIST Quarterly 37. Available: http://www.iassistdata.org/downloads/iqvol371_4_altman.pdf. Accessed: 23 September 2015.
4. Bandrowski A, Brush M, Grethe JS, Haendel MA, Kennedy DN, Hill S, et al. (2015) The Resource Identification Initiative: A cultural shift in publishing [version 1; referees: 2 approved] F1000Research 2015, 4:134. doi: 10.12688/f1000research.6555.1
5. The FAIR data Guiding Principles [Internet]. Available: https://www.force11.org/group/fairgroup/fairprinciples. Accessed: 23 September 2015.
6. Starr J, Castro E, Crosas M, Dumontier M, Downs RR, Duerr R, et al. (2015) Achieving human and machine accessibility of cited data in scholarly publications. PeerJ PrePrints 3:e697v4. https://dx.doi.org/10.7287/peerj.preprints.697v4
7. Bradner S (1997) Key words for use in RFCs to Indicate Requirement Levels. [Internet]. Available: http://www.ietf.org/rfc/rfc2119.txt. Accessed: 23 September 2015.
8. Local Resource Identifier Scheme. [Internet] Available: http://purl.org/spar/datacite/local-resource-identifier-scheme. Accessed: 23 September 2015.
9. Zeeberg BR, Riss J, Kane DW, Bussey KJ, Uchio E, Linehan WM, et al. (2004) Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics. BMC Bioinformatics. 23: 80.
10. Juty N, Le Novère N, Hermjakob H, and Laibe C. (2012) Delivering Cool URIs that Don't Change. Proceedings of the Semantic Web Applications and Tools for the Life Sciences (SWAT4LS) 2012. http://ceur-ws.org/Vol-952/paper_4.pdf.
11. Sporny M (2011) The Case for Curies. [Internet] Available: http://manu.sporny.org/2011/case-for-curies/. Accessed: 23 September 2015.
12. W3C Working Group (2010) CURIE Syntax 1.0. [Internet] Available: http://www.w3.org/TR/curie/. Accessed: 23 September 2015.
13. Kunze J, Rodgers R. ARK Specification [Internet] Available: http://www.cdlib.org/services/uc3/arkspec.pdf. Accessed: 23 September 2015.
14. Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek BE, et al. (2009) Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinformatics. 10: 136. doi: 10.1186/1471-2105-10-136.
15. Kratz J, Strasser C. Data publication consensus and controversies [v3; ref status: indexed, http://f1000r.es/4ja] F1000Research 2014, 3:94 doi: 10.12688/f1000research.3979.3
16. Gray AJG, Baran J, Marshall MS, Dumontier M. (2014) Identifiers in Dataset Descriptions: HCLS Community Profile. In: HCLS Community Profile [Internet]. W3C 2014 - . Available: https://htmlpreview.github.io/?https://github.com/indiedotkim/HCLSDatasetDescriptions/blob/master/Overview.html#s6_3. Accessed: 23 September 2015.
17. Complete Example of a Dataset Description http://www.w3.org/TR/hcls-dataset/#appendix_1. [Internet] Available: http://www.w3.org/TR/hcls-dataset/#appendix_1 Accessed: 23 September 2015.
18. DIPPER: The Monarch Data Ingest Pipeline. [Internet] Available: https://github.com/monarch-initiative/dipper/blob/master/README.md#identifiers Accessed: 23 September 2015.
19. Mietchen D, McEntyre J, Beck J, et al. (2015) Force11 Data Citation Implementation Group Adapting JATS to support data citation. In: Journal Article Tag Suite Conference (JATS-Con) Proceedings 2015 [Internet] Available: http://www.ncbi.nlm.nih.gov/books/NBK280240/?report=classic. Accessed: 23 September 2015.
20. Klein M, Van de Sompel H, Sanderson R, Shankar H, Balakireva L, Zhou K, et al. (2014) Scholarly context not found: one in five articles suffers from reference rot. PLoS One 9: e115253. doi: 10.1371/journal.pone.0115253. eCollection 2014