10 Simple rules for design, provision, and reuse of persistent identifiers for life science data

Julie A McMurry[1]* , Niklas Blomberg[2], Tony Burdett[1], Nathalie Conte[1], Michel Dumontier[3], Donal Fellows[4], Alejandra Gonzalez-Beltran[5], Philipp Gormanns[6], Janna Hastings[1], Melissa A Haendel[7], Henning Hermjakob[1], Jean-Karim Hériché[8], Jon C Ison[9], Rafael C Jimenez[2], Simon Jupp[1], Nick Juty[1], Camille Laibe[1], Nicolas Le Novère[1,10], James Malone[1], Maria Jesus Martin[1], Johanna R McEntyre[1], Chris Morris[11], Juha Muilu[12], Wolfgang Müller[13], Christopher J Mungall[14], Philippe Rocca-Serra[5], Susanna-Assunta Sansone[5], Murat Sariyar[15,16], Jacky L Snoep[17,18], Natalie J Stanford[4], Neil Swainston[19], Nicole L Washington[14], Alan R Williams[4], Katherine Wolstencroft[20], Carole Goble[4], Helen Parkinson[1]

1 European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom
2 ELIXIR Hub, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom
3 Center for Biomedical Informatics Research, Stanford University, Stanford, California, USA
4 School of Computer Science, The University of Manchester, Manchester, United Kingdom
5 Oxford e-Research Centre, University of Oxford, Oxford, United Kingdom
6 Institute of Experimental Genetics, Helmholtz Centre Munich -German Research Center for Environmental Health (GmbH), Neuherberg, Germany
7 Department of Medical Informatics and Epidemiology and OHSU Library, Oregon Health & Science University, Portland, USA.
8 European Molecular Biology Laboratory, Heidelberg, Germany
9 Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark
10 Babraham Institute, Cambridge, United Kingdom
11 STFC, Daresbury Laboratory, Warrington, United Kingdom
12 Genomics Coordination Center, Department of Genetics, University Medical Center Groningen and Groningen Bioinformatics Center, University of Groningen, Groningen, Netherlands
13 SDBV, HITS, Heidelberg, Germany
14 Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
15 Institute of Pathology, Charite – University Medicine Berlin, Berlin, Germany
16 TMF – Technologie- und Methodenplattform e. V. Berlin, Germany
17 MIB, University of Manchester, Manchester, UK
18 Department of Biochemistry, Stellenbosch University, Stellenbosch, South Africa
19 Manchester Centre for Synthetic Biology of Fine and Speciality Chemicals (SYNBIOCHEM), University of Manchester, Manchester, UK.
20 Leiden Institute of Advanced Computer Science, Leiden University, Leiden, Netherlands

* Corresponding authors
E-mail: mcmurry.julie@gmail.com, parkinson@ebi.ac.uk

1. Keywords: Identifiers, Identifier design, Accessions, Databases, Big data, Interoperability, PID, Data citation, Synthesis research, Reproducibility, Standards, e-Science

Introduction

When we interact, we use names to identify things. Usually this works well, but there are many familiar pitfalls. For example, the “morning star” and “evening star” are both names for the planet Venus. “The luminiferous ether” is a name for an entity which no one still thinks exists. There are many women named “Margaret”, some of whom go by “Maggie” and some of whom have changed their surnames. We use everyday conversational mechanisms to work around these problems successfully. Naming problems have plagued the life sciences since Linnaeus pondered the Norway spruce; in the much larger conversation that underlies the life sciences, problems with identifiers (**Box 1**) impede the flow and integrity of information. This is especially challenging within “synthesis research” disciplines such as systems biology, translational medicine, and ecology. Implementation-driven initiatives such as ELIXIR, BD2K, and others (**Text S1**) have therefore been actively working to understand and address underlying problems with identifiers.

Good, global-scale, persistent identifier design is harder than it appears, and is essential for data to be Findable, Accessible, Interoperable, and Reusable (Data FAIRport principles [1]). Digital entities (e.g., files), physical entities (e.g., biosamples), and descriptive entities (e.g., ‘mitosis’) have different requirements for identifiers. Identifiers are further complicated by imprecise terminology and different forms (**Box 1**).

Of the identifier forms, `Local Resource Identifiers (LRI)` and their corresponding `full Uniform Resource Identifiers (URIs)` are still among the most commonly used and most problematic identifiers in the bio-data ecosystem. Other forms of identifiers such as `Uniform Resource Name (URNs)` are less impactful because of their current lack of uptake. Here, we build on emerging conventions and existing general recommendations [2,3] and summarise the identifier characteristics most important to optimising the flow and integrity of life-science data (**Table 1**). We propose actions to take in the identifier ‘green field’ and offer guidance for using real-world identifiers from diverse sources.

Box 1. Identifier forms and terminology

An **`Identifier`** is a sequence of characters that identifies an entity

- **`Local Resource Identifier (LRI)`** is an identifier that is only guaranteed to be unique within a single database
    - Databases and library systems often refer to the Local Resource Identifier as an ‘Accession’. In web architecture, the LRI is sometimes referred to simply as ‘web resource’.
    - LRI formats vary by provider and may have subparts; subparts are not described here as they are non uniform. For example, a LRI may be opaque (eg. A0A022YWF9) or recognizable (ZDB-GENE-980526-388)
- **`Uniform Resource Identifier (URI)`** is an identifier that is guaranteed to be both uniform and globally unique
    - **`CURIE`** is a compact URI comprised of **`<Prefix>:<LRI>`** wherein:
        - **`Prefix`** is:
            - a mnemonic that helps in human communication
            - documented and aspirationally globally unique
            - documented in terms of its case convention
            - conforms to the rules of an XML QName (e.g. does not contain ‘:’), and
            - deterministically expandable to **`a resolving namespace`** (see below) which is the basis for the CURIE’s global uniqueness.
    - **`full URI`** is an ASCII string that uniquely identifies a resource and also resolves to (provides or redirects to) a webpage containing information about the identified entity. **`Full URIs`** are generally HTTP but may be other (eg. HTTPS).

A **`resolving namespace`** is a sequence of characters which, when prepended to the LRI, yields the **`full URI.`** Occasionally, the **`resolving namespace`** is the same as the homepage eg. http://zfin.org/ in **Fig. 1**.
In all cases, the **`resolving namespace`** must be exactly as it appears in the full URI: it must include the protocol (eg. http://) and, if applicable, trailing slash or other delimiters.
See also Figure 1 examples and supplementary glossary (Table S2) for additional terms and concepts.

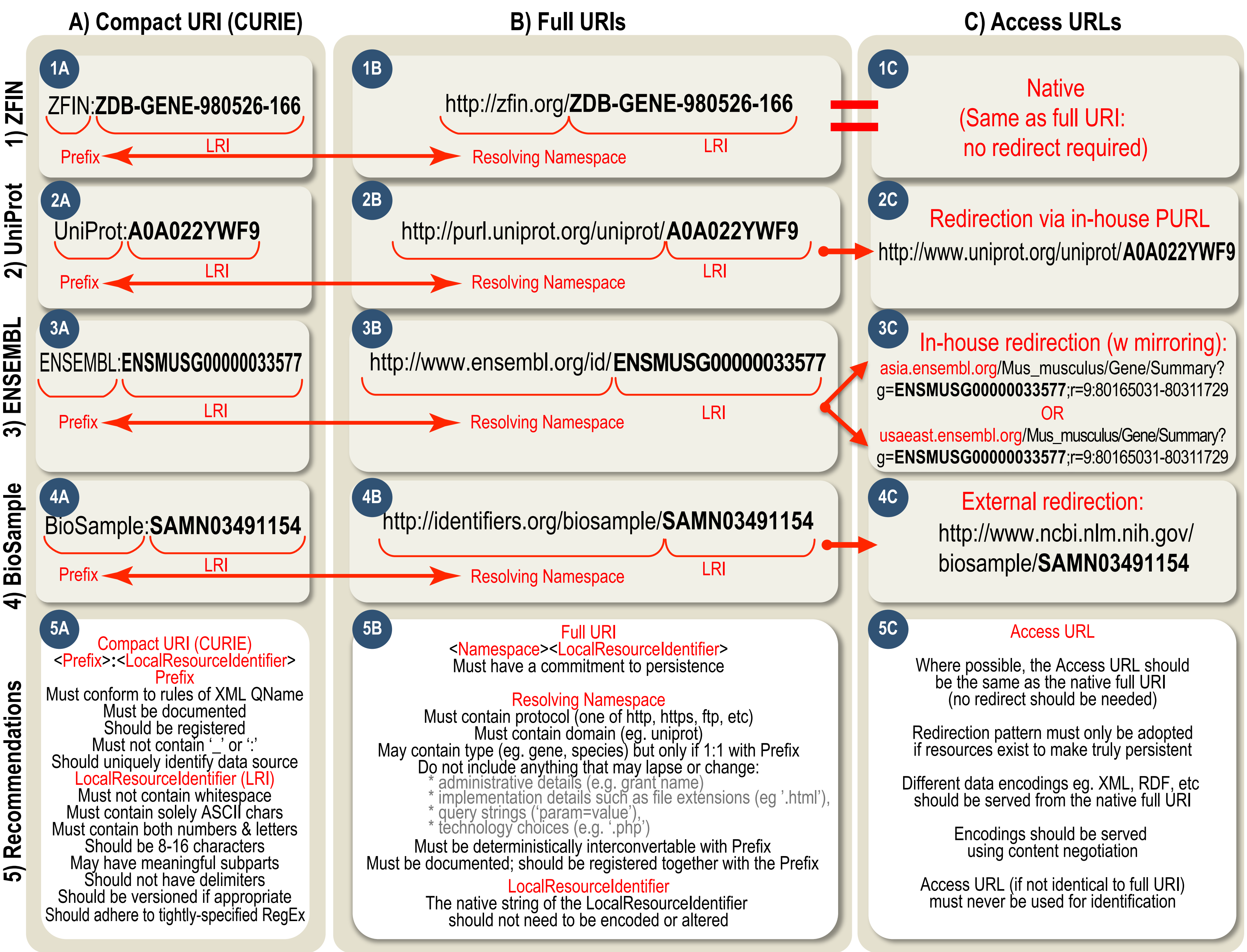
A) Compact URI (CURIE)
B) Full URIs
C) Access URLs
1) ZFIN
2) UniProt
3) ENSEMBL
4) BioSample
5) Recommendations
1A
ZFIN:ZDB-GENE-980526-166
Prefix
LRI
1B
http://zfin.org/ZDB-GENE-980526-166
Resolving Namespace
LRI
1C
Native
(Same as full URI:
no redirect required)
2A
UniProt:A0A022YWF9
Prefix
LRI
2B
http://purl.uniprot.org/uniprot/A0A022YWF9
Resolving Namespace
LRI
2C
Redirection via in-house PURL
http://www.uniprot.org/uniprot/A0A022YWF9
3A
ENSEMBL:ENSMUSG00000033577
Prefix
LRI
3B
http://www.ensembl.org/id/ENSMUSG00000033577
Resolving Namespace
LRI
3C
In-house redirection (w mirroring):
asia.ensembl.org/Mus_musculus/Gene/Summary?
g=ENSMUSG00000033577;r=9:80165031-80311729
OR
usaeast.ensembl.org/Mus_musculus/Gene/Summary?
g=ENSMUSG00000033577;r=9:80165031-80311729
4A
BioSample:SAMN03491154
Prefix
LRI
4B
http://identifiers.org/biosample/SAMN03491154
Resolving Namespace
LRI
4C
External redirection:
http://www.ncbi.nlm.nih.gov/
biosample/SAMN03491154
5A
Compact URI (CURIE)
<Prefix>:<LocalResourceIdentifier>
Prefix
Must conform to rules of XML QName
Must be documented
Should be registered
Must not contain '_' or ':'
Should uniquely identify data source
LocalResourceIdentifier (LRI)
Must not contain whitespace
Must contain solely ASCII chars
Must contain both numbers & letters
Should be 8-16 characters
May have meaningful subparts
Should not have delimiters
Should be versioned if appropriate
Should adhere to tightly-specified RegEx
5B
Full URI
<Namespace><LocalResourceIdentifier>
Must have a commitment to persistence
Resolving Namespace
Must contain protocol (one of http, https, ftp, etc)
Must contain domain (eg. uniprot)
May contain type (eg. gene, species) but only if 1:1 with Prefix
Do not include anything that may lapse or change:
* administrative details (e.g. grant name)
* implementation details such as file extensions (eg '.html'),
* query strings ('param=value'),
* technology choices (e.g. '.php')
Must be deterministically interconvertable with Prefix
Must be documented; should be registered together with the Prefix
LocalResourceIdentifier
The native string of the LocalResourceIdentifier
should not need to be encoded or altered
5C
Access URL
Where possible, the Access URL should
be the same as the native full URI
(no redirect should be needed)
Redirection pattern must only be adopted
if resources exist to make truly persistent
Different data encodings eg. XML, RDF, etc
should be served from the native full URI
Encodings should be served
using content negotiation
Access URL (if not identical to full URI)
must never be used for identification

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

Characteristics	Definition	Recommendation
Unambiguous	One identifier is associated to no more than one entity	LRI must locally full URI must globally
Unique	One entity is identified by no more than one identifier	LRI & URI should
Stable (identifier)	The identifier *itself* stays the same over time	LRI & URI must[a]
Stable (entity)	The identifier identifies the same entry (representation) over time (elements such as metadata may be refined without changing the entity nature)	LRI & URI must
Versioned	The identifier is versioned to reflect changes in the definition of the entity, or major changes in its descriptive metadata	LRI & URI should
Persistent	The representational association persists between the identifier and its identified entity	LRI & URI must
Web-resolvable	The identifier can be resolved to a web address where the data or information about the entry can be accessed	LRI & URI must
Defined	The identifier adheres to a formal pattern (e.g. regular expression)	LRI & URI must
Web-friendly	The `LRI` is of a format that does not need special handling when used in URLs and common exchange formats (e.g. XML)	LRI & URI must
Convertible	The `identifiers` can be inter-converted algorithmically	LRI & URI must
Free to assign	The identifier can be assigned at no cost to those depositing data in a repository	LRI & URI should
Open access and use	The identifier itself may be transparently referenced, and actioned (eg. in a public index or search) anywhere by anyone and for any reason. Restrictions on access or use of associated metadata or entities may apply but are not recommended.	LRI & URI should
Documented	The identifier scheme is documented	LRI & URI must

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

Rule 1: If you create identifiers, do not DIY (Do Identifiers by Yourself)

If you are creating new, identifiable knowledge, deposit it in suitable repositories whenever possible (**Table S4)**. In the absence of a suitable repository, or for any of the reasons below, if you must create your own identifiers, keep in mind that they will make their way into the broader data 'ecosystem'. So consider the scope of the entities to be identified, the way that data is/will be generated and how it will be consumed, as well as existing identifier platforms and services [4]. Determine your identifier-management strategies before creating any identifiers: all approaches require planning and coordination at some stage (**Table S4**).

When people create alternate identifiers, it is usually to a) reduce risks posed by dependency on an outside source or b) identify meaningful differences in an entity, its state, or its representation. Whatever the case, if you must create your own identifier, you must also clearly characterize the relationship to the existing identifier (eg. 'derived from', 'related to'). In the case of factual corrections, 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: Help identifiers travel well: don't let them leave home without a Prefix and a 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 1**) alone is not up to these tasks because of inevitable collisions. For instance, the LRI "9606" corresponds to at least six different entities[5]: a Pubmed article, a CGNC gene, a PubChem chemical, as well as an NCBI taxon, a BOLD taxon, and a GRIN taxon. Full URIs (**Box 1**) are the only identifier appropriate for machine-driven global data integration tasks; however their length makes them unwieldy for humans working with the data or for referencing in publications or other text. The full URIs should be presented in CURIE form (**Box 1**) to human users/readers when identifiers are referenced outside of their native context. CURIEs are location-documented but not location-dependent; this makes them useful in answering questions about resources that exist in more than one location (eg. "How do I locate X within portal Y?").

Therefore, if you are a database provider, document your preferred Prefix (**Box 1**) and its binding to a resolving `namespace` (**Box 1**). It is in your best interests to make reasonable efforts to avoid prefix collisions, especially where the corresponding datasets are likely to be used in the same context. We strongly recommend that you record your `prefix` and `resolving namespace` in the appropriate registry/registries (**Table S3**).

Rule 3: Make Local Resource Identifiers rugged to real-world use

Pre-existing identifiers should be reused without modifications (see Rule 10); however, in a green field, there are a few design choices that can help new LRIs perform better beyond their local scope:

- Must comprise only printable ASCII characters without whitespace; this guards against corruption and mistranscription.
- Should not contain ':', a reserved character for parsing in CURIEs **(Box 1)**
- Should not contain '.' except to denote version (see Rule 7)
- If additional delimiters (other than ':' and '.') are needed, prefer '-'
- Should contain both letters and numbers to avoid misinterpretation as numeric data (eg. truncation of leading zeros)
- Should not be a pattern that could result in misinterpretation whether as dates, exponents [6], or unintended words (fictional examples of problem LRIs would be "5e1234" or "may-15")
- Case convention must be fixed and documented, but should be case insensitive
- Must adhere to a formal pattern (regular expression); this adherence facilitates, but does not guarantee, validation and retrieval from scientific text. To minimize awkwardness in prose, 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 LRI collisions, it is considerate to tightly specify your pattern (eg. using two or more fixed letters at the start of the string).

Rule 4: Make the full URI simple and durable

If you are a database provider, you must implement full URIs (**Fig. 1 panel B**) for your outward-facing identifiers to be "resolvable" to a web page. Full URIs must be deterministically interconvertible with the `Compact URIs` (**Fig. 1 panel A**). Full URIs in turn must resolve to a "landing page" ("access URL", **Fig. 1 panel C**). If you have the resources to support your own full URIs that are truly persistent, design them to be as simple as possible. 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`'). Specific recommendations are summarised in the lower panel of **Fig. 1 panel B**. However, if long-term persistence of your native full URI is in question, you must use a dedicated suitable resolver service. When choosing a resolver, use one that is JDDCP-adherent [4] and be mindful of any constraints you may have (**Text S5**). Whether or not you outsource resolution to a service, implement best practice [4] on serving landing pages and different encodings of your data (aka "content negotiation").

Fig. 1. Examples and recommendations for identifiers in different forms: Compact URIs (CURIEs) (Panel A), full URIs (Panel B) and Access URLs (Panel C), each of which has corresponding examples (ZFin, UniProt, ENSEMBL, BioSample) followed by a summary of recommendations for new identifier designs. In each case the LRI adheres to Rule 3, the full URI can be algorithmically derived from the CURIE, and the LRI itself is included (unmodified) within the full URI.

Rule 5: Carefully consider whether to embed meaning

When designing new identifiers, be explicit about what it is they identify, but carefully consider how to convey this meaning--whether embedded in the identifier itself, or in the metadata. Meaning is never required to be embedded in an identifier; for instance, UniProt LRIs are meaning-free but nevertheless adhere to Rule 3. Meaning may be embedded where 1) durable, 2) coarse-grained, 3) uncontested, and also 4) useful to the data consumer, but only if all four conditions apply without potential edge cases.

Except where durable and deterministic (e.g. an InChI string identifying a chemical structure), you should not embed information that is at the per-entity level (such as name or label). Never embed its type if an entity could change from one type to another, for example if the type depends on the entity's developmental stage or if the typing nomenclature is not well defined scientifically. If a database name may change, it should not be embedded. These rules of thumb apply especially to LRIs but also to the path of the full URIs. Keep in mind that each prefix must correspond 1:1 with a `resolving namespace`. If possible, avoid varying URI paths by entity type, authority, etc... as this can be confusing for users.

Rule 6: Make the full URI and CURIE clear and easy to find

Make full URIs as obvious as possible to users, especially where these may differ from access URLs or application pages. For instance, at the record-level, advertise the "permanent link" together with a statement about persistence (see for instance http://ensembl.org/id/ENSMUSG00000033577). Ideally, the permanent link to the most recent version should be provided as well. Although it is good for a database provider to include general documentation regarding citation, it is even better to also provide a "cite this" button at the level of the resource page.

If source LRIs already have a colon, database providers must make it clear to users what the corresponding CURIE syntax is. We recommend referencing it as if it were *already* a CURIE. For instance, the case of GO:0007049, the prefix 'GO' can be expanded to http://purl.obolibrary.org/obo/GO_ and prepended to the numeric fragment to yield http://purl.obolibrary.org/obo/GO_0007049 in accordance with their documentation. For DOIs, the citation convention is that the `prefix` (as defined in this writing) is "doi"; the corresponding `namespace` would be https://dx.doi.org/ and the LRI everything that follows. If the provider chooses a different resolver, the provider's prefix (e.g. "BioSample" **Fig. 1, panel 2A**) must expand to a `resolving namespace` which is the concatenation of resolver and provider (e.g. http://identifiers.org/biosample/, **Fig. 1, Panel 2B**).

Rule 7: Implement a version-management policy

Changes in data resources impact how they can be referenced and used. If you issue identifiers, either document the change history for the resource (see also Rule 8), or version the identifier itself, or do both. Whatever the approach, it must be clearly documented. 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 changes are "non-breaking", it is reasonable to instead maintain a machine-actionable change history wherein the changes are also meaningfully categorized. Versioning and change history work well in combination, especially when multiple types of changes overlap. Even when previous records are archived or removed, the full URI should continue to resolve, but to a "tombstone" page. A summary of versioning recommendations follows in Table 2 below, with UniProt [7] identifiers as examples. See Kratz et al. [8] for a more in-depth discussion of change management considerations.

Table 2. Recommendation for versioning with full URIs

Behavior	Level	Example (for clarity, LRI only is shown)
Add version information after a dot	Should[a]	P12345.3
Base resource resolves (302 redirect) to the most recent version	Must	P12345
Base resource deterministically convertible from version	Should	P12345.1 to P12345
Older versions resolve	Must	P12345.1
Illegal or invalid version produces informative error message	Must	P12345.302
Link from older version to current version is provided	Must	P12345.1
A list of all previous versions is available	Should	P12345 ('history' linked)
Two versions (or dates) can be compared	Should	http://www.uniprot.org/uniprot/P12345?version=*

[a]If versioning at the individual record level (eg. UniProt), you must version after the dot; this enables a single CURIE prefix to be used. If versioning a whole database, you may version in the namespace (eg. Ensembl).

Rule 8: Manage complex lifecycles without deletion

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 remain mainly static once generated, concept descriptions evolve rapidly; even the nature and number of the relevant metadata fields changes 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
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 the primary identifier 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 resolve and should provide a warning and pointers to the new ones.	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
Obsoletion: If an entry has been removed or deprecated, the original identifier must still resolve. 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 ('replaced_by') or suggested ('consider'). The obsoleted ID must never be reassigned to another entity. A list of obsoleted IDs should be maintained.	http://www.uniprot.org/uniprot/A0AV18

Rule 9: Document the identifiers you issue and use

A healthy global-scale identification cycle is a shared responsibility and provider/consumer roles often overlap. Whether you issue identifiers, or just reference the identifiers of others, document how your IDs are assigned and managed. These 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 Health Care and Life Sciences Interest Group [9]; the format of the description may vary. **Table 4** provides a set of questions that can be used to develop such documentation.

Table 4. Questions that good identifier documentation should answer

Scope	Question to answer	Recommendation
Provider	What is your preferred `Prefix`? If it is registered, where? What is the CURIE?	Must include
Provider	What is your primary resolution namespace, if only one exists? If multiple, equally-valid resolution namespaces co-exist, what are these? e.g. INSDC.org has four such schemes as the entire dataset is fully represented by each of four authorities: NCBI, GenBank, ENA, and DDBJ	Must include
Provider	What alternate resolution namespaces, if any, are known to have been used by others? (Even though alternates are not recommended for use, knowing what these schemes are facilitates data integration.)	Should include
Provider	What is the persistence policy regarding maintenance of the full URIs? For corresponding entities/metadata?	Must include
Provider	Can machine-readable representations of your entities be accessed? If so, where and in what formats?	Must include
Provider	What is the regular expression of the LRI?	Strongly recommended
Provider	What types of entities are identified, what is the scope of these entities?*	Should include
Provider	Are there relationships between identifiers? Where are these described?*	Should include
Provider	Under what license are identifiers made available?	Should include
Provider	Does the lifecycle of the entities potentially include versioning, splitting, merging, or deprecation? How are these changes managed, communicated, and synchronised between those using that entity?*	Must include
Provider	Do you identify *entities* that are also identified by others? Who are these others? Where are these mappings found and who, if anyone, maintains them?	Strongly recommended
Provider-User	Do you reference *identifiers* that are issued by other authorities? If so, in what cases? How often are the identifiers synchronised?	Must include
Provider-User	If you reference *identifiers* that are issued by other authorities, what are the prefix-to-resolving-namespace mappings used? What is the source of these mappings (eg. manual or identifier service). Where can your mappings be found?	Must include

* Adapted from the Linked open data institute recommendations [10]

Rule 10: Reference responsibly and rely on full URIs

When provider responsibilities (Rules 1-9) are met, the corresponding consumer responsibilities are straightforward (**Table S6**). In practice, data consumers work in the real world of identifiers from heterogeneous sources: When publishing a dataset or database with external database references:

- You must document and maintain your prefix-to-namespace bindings (see details in Rule 9) and do so in a machine-readable format.
- You should defer to provider regarding their preferred prefix and resolving namespace, or at a minimum use an identifier service namespace.
- In cases of undocumented prefixes or URIs, you should defer to the data provider or, if undocumented, look them up in an identifier service.
- In cases of a prefix collision (e.g., the prefix 'GEO' refers to Gene Expression Omnibus as well as the GeoNames Ontology), you should ideally defer to the respective data providers about how to modify the prefixes.
- If a reference points to an identifier that no longer resolves, you should contact the provider. At a minimum, your annotation should be modified to reflect the deletion.
- "Official" full URIs may not be documented or adhered to. Services such as sameas.org, myEquivalents [11] as well as CURIEs can be used to help find potential co-references between different data sets.

Conclusion

Better identifier design, provisioning, documentation, and referencing can go a long way to address many of the identifier problems currently faced in the life science data cycle. We recognize that improved software tooling for identifiers would lower barriers to adoption of these rules. We also recognize the need for formal software engineering specifications of identifier formats (e.g., regular expressions, Backus Naur Form), and/or alignment between existing specifications, and hope that this paper can catalyze such efforts.

Acknowledgments

This work was supported primarily 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, T Burdett, and J Malone. ELIXIR core funds supported N Blomberg, R Jimenez. The European Commission provided additional support for Simon Jupp under grant number 601043 ("DIACRHON") 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.

Additional support from the BBSRC was received:

- 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") for N Swainson
- 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

Additional support from the NIH was received:

- 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, 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, and Eleanor Williams for their valuable feedback and suggestions. We also wish to thank Lee Harland for the suggestion to address this important issue.

References

1. JDDCP (Joint Declaration of Data Citation Principles). The FAIR data Guiding Principles [Internet]. JDDCP. 2011 Oct - [cited 2015 May 15]. Available: https://www.force11.org/group/fairgroup/fairprinciples
2. Juty N, Le Novère N, Laibe C. Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. Nucleic Acids Res. 2012 January. doi: 10.1093/nar/gkr1097
3. Identifier and Metadata Working Group (DCIG). Identifier Best Practices. 2014 June 14. [Cited 2015 May 15]. [Internet]. Available: http://www.scc.lancs.ac.uk/research/projects/researchobject/mediawiki-1.22.6/index.php/Identifier_best_practices
4. 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
5. McMurry, JA. (2015). Identification at local and global scale: a case for using the Compact URI (CURIE) for life science data. Zenodo. 10.5281/zenodo.17576
6. Zeeberg BR, Riss J, Kane DW, Bussey KJ, Uchio E, Linehan WM, et al. Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics. BMC Bioinformatics. 2004 Jun 23;5:80.

7. Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek BE, et al. Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinformatics. 2009 May 8;10:136. doi: 10.1186/1471-2105-10-136.
8. Kratz J and 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)
9. Gray AJG, Baran J, Marshall MS, Dumontier M. Identifiers in Dataset Descriptions: HCLS Community Profile. 2014 [cited 15 May 2015]. In: HCLS Community Profile [Internet]. W3C 2014 - . Available: https://htmlpreview.github.io/?https://github.com/indiedotkim/HCLSDatasetDescriptions/blob/master/Overview.html#s6_3
10. Open Data Institute and Thomson Reuters, 2014, Creating Value with Identifiers in an Open Data World, retrieved from thomsonreuters.com/site/data-identifiers/
11. My Equivalents [Computer software]. [Cited 2015 May 15]. Available: https://github.com/EBIBioSamples/myequivalents/wiki. doi: 10.5281/zenodo.17555