Human-readable and machine-readable Persistent Identifiers

Martin Fenner, Gobbledygook

May 27, 2015

Yesterday Julie McMurry and co-authors published a preprint **10 Simple rules for design, provision, and reuse of persistent identifiers for life science data** (McMurry et al., 2015). This is an important paper trying to address a fundamental problem: how can we make persistent identifiers both human-readable and machine-readable?

Don't be fooled by the title (used frequently by PLOS Computational Biology) - the paper doesn't describe simple rules that help the average life sciences researcher. Rather, the paper deals with rather complex issues, and has 36 authors.

There is general agreement that we need persistent identifiers for scholarly communication, and that also includes life sciences datasets, the focus of the paper. What is less clear is how to express these persistent identifiers. An identifier such as **AB020317** - for the mouse p53 gene - is ambiguous. It is not clear without additional information that this is an identifier for the GenBank nucleotide database, rather than something completely different. One common approach to make this identifier unambiguous is to use URIs (Uniform Resource Identifiers), e.g. http://www.ncbi.nlm.nih.gov/nuccore/AB020317 in this case.

The paper doesn't like this approach, and even states that "URIs are still among the most commonly used and most problematic identifiers in the bio-data ecosystem". The text also states that "their length makes them unwieldy for humans working with the data or for referencing in publications or other text", but doesn't go into any detail why URIs are "problematic identifiers", or why length is an issue in an online environment.

This is an important weakness of the paper, because the authors propose an alternative: CURIEs or **compact URIs**. CURIEs were proposed by the W3C a few years ago, as a way to make URIs more human-readable. The idea is simple, we use a namespace in addition to the local identifier, separated by a colon, e.g. **Genbank:AB020317**.

This approach has of course been common practice in the life sciences before CURIEs or even the WWW existed, and is still the most common approach how

identifiers for life sciences data are referenced in the scholarly literature. Unfortunately there are important problems with CURIEs, most of them mentioned in the paper:

- Persistent identifiers need to be resolvable, without additional information we don't know what to do with **Genbank:AB020317**. Most life sciences researchers understand this CURIE, but that might not necessarily be true for less commonly used namespaces
- Namespaces are not necessarily unique, the paper uses **GEO** (which could mean Gene Expression Omnibus or GeoNames Ontology) as an example
- Rule 3 in the paper goes into great detail what characters and patterns should be avoided in local identifiers that are part of a CURIE. It is not clear whether these recommendations will always be followed or how to check them
- CURIEs should follow a pattern (regular expression) so that they can be extracted from a text. We know (Kafkas, Kim, & McEntyre, 2013) that extracting identifiers from journal articles is possible, but difficult

URIs don't have the problems listed above: they resolve, are unique, and there is good understanding (and available tools) of how a valid URI should look like and how to extract URIs from text documents. That is why URIs are good representations of persistent identifiers.

Another problem I have with CURIEs: the idea doesn't seem to have caught on from the initial work more than five years ago (background reading here). I'm not even sure what percentage of persistent identifier experts know about CURIEs.

My recommendation for life sciences data: express persistent identifiers as URIs. Now that can go into 10 simple rules for the average life sciences researcher.

P.S. This blog uses a tool I wrote two years ago that automatically turns CURIEs in the text into links.

References

Kafkas, Ş., Kim, J.-H., & McEntyre, J. R. (2013). Database Citation in Full Text Biomedical Articles. *PLoS ONE*. doi:10.1371/journal.pone.0063184

McMurry, J., Blomberg, N., Burdett, T., Conte, N., Dumontier, M., Fellows, D. K., ... Parkinson, H. (2015). 10 Simple rules for design, provision, and reuse of persistent identifiers for life science data. doi:10.5281/zenodo.18003