Ten simple rules for biomedical ontology development

Mélanie Courtot
EMBL-EBI
Hinxton, UK
mcourtot@gmail.com

James Malone
FactBio Ltd
Cambridge, UK
james@factbio.com

Christopher J Mungall
Lawrence Berkeley National Laboratory
Berkeley, USA
cjmungall@lbl.gov

Abstract—**Biomedical ontology development is often a time and resource consuming endeavor. To maximize efficiency of the process, we present a set of 10 simple rules covering basic technical requirements such as scoping and versioning, while considering additional elements such as licensing and community engagement. When applied, the rules will help avoid common pitfalls and jump-start ontology building.**

Keywords—ontology development, tutorial, rules

INTRODUCTION

Biomedical Ontologies are notoriously challenging and laborious to develop, despite their uncontested usefulness for data description, sharing and integration. As the amount of data generated keeps increasing, ontologies are becoming a de facto requirement for scientific creation and maintenance of datasets. While advantages to using an ontology are many, it is not straightforward for inexperienced users to choose which to use [1] before considering development of their own. Additionally, there is often no single resource providing exactly what is needed, and many biologists embark on a new ontology building task without being fully aware of some basic notions in ontology development. This paper seeks at documenting some general rules and guide neophyte users towards practical considerations for efficient biomedical ontology building.

I. SET THE SCOPE FROM USERS' NEEDS

It is often very tempting to 'dig in', and start creating new terms and organize them in a hierarchy. However before proceeding with ontology development itself, it is crucial to take a step back and consider the use cases the ontology is attempting to address. Typically this is in the form of *competency questions* [2] - queries which the ontology should be able to satisfy in order to be considered correct and usable. Building an ontology from the bottom up will ensure there is coverage, i.e. the 'terms' required are present, but it will not always ensure that the queries required are satisfiable. This requires an understanding of those questions and from there building in class descriptions and structure such that they can be answered by the ontology.

II. DO YOUR RESEARCH & REUSE AS MUCH AS POSSIBLE

When choosing to create a new resource, care should be taken to reuse work done in the context of other efforts where possible. While this introduces additional constrains such as need to keep in sync or decisions about positioning and modifications, the advantages of doing so greatly outweigh the disadvantages. Reusing terms from other resources allows developers to rely on the knowledge of domain experts who curated them and to dedicate more work time for novel terms. The Minimum Information to Reference an External Ontology Term guidelines [3] specifies a mechanism to selectively import a term from a source ontology into a target resource, without the overhead of importing the whole external file. For example, the Gene Ontology (GO, [4]) currently imports selected terms from the Chemical Entities of Biological Interest (ChEBI, [5]) to model physiological responses to drugs. Avoiding duplication of resources additionally increases interoperability: a single URI is created per term, preventing the need for tedious mappings between terms with the same meaning in different resources.

III. PUBLISH THE ONTOLOGY LICENSE AND ATTRIBUTION MODEL

When building an ontology, you should think about licensing early on [6]. Indeed, licenses cannot be made more restrictive; they can only be loosened towards a more permissive one. Within the OBO Foundry [7], we chose to recommend the Creative Commons licenses [8], specifically CC-by, which requires attribution upon reuse. The OBO Foundry only requires the original URIs be reused for attribution, which prevents 'attribution stacking' : only the URI need to be cited, without the need for adding extra citations to individuals or projects. However other efforts such as Wikidata [9] require resources be available under CC-0 (i.e. public domain) for reuse, so the chosen license can and will impact the usage that can be made of your resource. Proper attribution will be important when trying to track usage, and can help justify supporting it to funding agencies.

IV. PROVIDE STABLE URIS & VERSION YOUR ONTOLOGY

While ontologies evolve through time, stability of identifiers is a fundamental tenet of their life cycle. Each entity described should have a unique identifier, and this identifier should be stable through time [10]. When terms become obsolete, a deprecation policy such as this of the GO [11] should be followed. Using URLs as identifiers enables for their dereferencing, i.e., resolution into human readable information in a browser as well as RDF for machine in the background.

The adoption of the OBO Foundry ID policy by many OBO library resources has enabled common tooling to be built, such as Ontobee [12] which provides built-in dereferencing for OBO resources.

V. Use a version control system or equivalent

Version Control Systems (VCS) allow for storage of ontologies and their versions in a common shared space, with a history of all edits preserved in a transparent fashion. In the world of software engineering, almost all software is developed using a VCS, and we argue that the same should hold for ontology engineering. In particular, we advocate for the use of a publicly hosted VCS system, such as GitHub or GitLab. These systems also provide mechanisms to help make stable releases, as well as provide issue trackers and tools to allow the wider community to interact with and comment on aspects of the development process. Many ontology developers have chosen to adopt a common folder structure with which to organize project file, which helps users find things in consistent places. Tools such as the ontology-starter-kit [13] can help you bootstrap a project using a standard layout.

VI. Use a common metadata set

Usage of common annotation properties allows tool developers to rely on them to build their user interface, and enables users to go back and check on the origin of the term and what its intended meaning is, and/or contact the relevant individual should they need more clarification about its usage. While it is usually non controversial that at least a label and definition be provided for each entity in the ontology, we found that other properties are useful in providing documentation and traceability. For example, source of the definition – such as a PMID or web citation - is often useful to capture and provides additional context for the term. Annotation properties should be used to indicate evolution of the ontology: 'replaced_by' indicates one-to-one replacement of obsolete terms and can be followed by scripts to update annotations for example, and 'creation_date' or 'created_by' can help audit the resource. A common metadata set [14] has been proposed and is currently used by many resources in the OBO Foundry. Other efforts exist to formalize metadata, such as the Simple Knowledge Organization System (SKOS) [15] and the Dublin Core (DC) Metadata element set [16].

VII. Evaluate early, often & openly

Collecting datasets first will ensure the resource developed fits the use case, and that there will be a gold standard against which the ontology can ultimately be evaluated. Some tools, such as the Ontology Lookup Service [17] allow calculating deltas (or diffs) between ontologies to explore their development, quality of content in terms of definitions, and compliance with ontology development best practice. For example, adherence to OBO Foundry principles [18] for ontology best design can provide *qualitative* evaluation. For external evaluation, other metrics can be useful to provide a *quantitative* overview, such as number of classes, properties in the ontology, or number of projects using the resource (as part of their own ontology or to annotate their datasets), evolution strategy, use of design patterns or domain interoperability. For either kind of evaluation, publish your results alongside with the ontology, pointing to the version that was being evaluated and changes that were being made when performing sequential evaluations.

VIII. Document your design patterns

Consider the knowledge you are trying to describe. In many cases in biology, a repetitive pattern can be seen. For example, the transport of a protein process in GO includes a starting point, an endpoint, a cargo, whether we are describing amino acid import into cell or oligopeptide export from mitochondrion. In the Ontology of Biomedical Investigations [19], assays are described via their input, output, and their evaluant (i.e., what is being measured). Using patterns for defining logical axioms allows for fast addition of new classes via script, as well as easier maintenance should the patterns be updated. Uberon documents a variety of anatomical entity design patterns on its wiki, and many of these are applicable to other ontologies [20]. The GO and several other ontologies including the Cell Type Ontology [21] already use standard patterns to generate new terms via the TermGenie tool [22]. In GO around 80% of new terms are added via this route. Other tools such as Tawny OWL [23] and the ontology Pre-Processing Language [24], for example as implemented via Webulous [25] are also available. A newer, simpler, version of templates is being implemented, 'Dead Simple OWL design patterns' (submitted). Adopting an upper level ontology can help ensure that the hierarchy developed is compliant with others which adhere to the same type of representation. This is important in the context of reuse of resources, or to ensure easy communication between developers. For example, 'cancer' can refer to a disease or an aggregate of cells, which would be in clearly separated areas of the ontology. Many upper ontologies are available [26]. In the OBO Foundry, the Basic Formal Ontology (BFO [27]) has been widely adopted.

IX. Make ontology as detailed as it needs to be. But no further.

Including users who understand the domain in question is also a valuable consideration. While there are many, freely available resources from which biomedical information can be collected, some are more reliable than others. Crowd sourcing such knowledge can be a productive method for collecting knowledge for inclusion into an ontology [9] but expertise from the biomedical domain in question is critical in ensuring the validity of the ontology content. Care should also be taken to capture the appropriate level of information. For example, when describing a disease, is only the diagnosis needed, or should the symptoms and signs be described as well? To maximize effectiveness, a resource need to abide by the Goldilocks principle [28] and capture *just the right amount* of information.

X. Engage with the community

Finally, don't be afraid to ask for help! There are many places where to get help, starting with the trackers of the resources you are interested in. The biomedical ontology community is relatively small, and many developers have been working together for a long time. While this means discussions can sometimes become heated, it also implies a long shared history and respect for each other's work. The community often comes together at yearly events such as the International Conference on Biomedical Ontology, the International Biocuration Conference or the Bio-ontology Special Interest Group. General mailing lists, such as public-semweb-lifesci@w3.org or obo-discuss@lists.sourceforge.net are also good places where to engage with other users and developers. Many other documents and blogs, such as Ontogenesis [29], can also provide assistance. Engaging a wider community means that in the longer-term more people may contribute, and will help establish a community of editors that provides some level of sustainability to the resource.

CONCLUSION

Building a new ontology can be a daunting task, and should not be taken on lightly. Good ontology development requires time and dedication, but if done correctly will provide advantages in storing and analysing biomedical data. Following a simple set of rules from early development on will prevent unnecessary proliferation of custom resources which are doomed to disappearing as their funding ends, and foster building of interoperable community resources.

ACKNOWLEDGMENTS

MC was funded by EMBL-EBI core funds. The authors would like to thank Helen Parkinson for helpful comments and suggestions on the manuscript.

REFERENCES

[1] James Malone, Robert Stevens, Simon Jupp, Tom Hancocks, Helen Parkinson, and Cath Brooksbank. Ten simple rules for selecting a bio-ontology. *PLOS Comput Biol*, 12(2):e1004743, 2016.

[2] Kamal Azzaoui, Edgar Jacoby, Stefan Senger, Emiliano Cuadrado Rodr´ıguez, Mabel Loza, Barbara Zdrazil, Marta Pinto, Antony J Williams, Victor de la Torre, Jordi Mestres, Manuel Pastor, Olivier Taboureau, Matthias Rarey, Christine Chichester, Steve Pettifer, Niklas Blomberg, Lee Harland, Bryn Williams- Jones, and Gerhard F Ecker. Scientific competency questions as the basis for semantically enriched open pharmacological space development. *Drug discovery today*, 18(17-18):843–52, sep 2013.

[3] M. Courtot, F. Gibson, A. L. Lister, J. Malone, D. Schober, R. R. Brinkman, and A. Ruttenberg. MIREOT: The minimum information to reference an external ontology term. *Applied Ontology*, 6(1):23–33, 2011.

[4] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology.TheGeneOntologyConsortium.*Naturegenetics*, 25(1):25–9, may 2000.

[5] Janna Hastings, Paula de Matos, Adriano Dekker, Marcus Ennis, Bhavana Harsha, Namrata Kale, Venkatesh Muthukrishnan, Gareth Owen, Steve Turner, Mark Williams, and Christoph Steinbeck. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic acids research*, 41(Database issue):D456–63, jan 2013.

[6] Science Commons - Ontology Copyright Licensing Considerations. Available from http:// sciencecommons.org/resources/readingroom/ ontology-copyright-licensing-considerations/, Accessed May 2016.

[7] OBO Foundry. OBO library. Available from http://obofoundry. org/, Accessed May 2016.

[8] Creative Commons. Creative commons licenses. Available from https: //creativecommons.org/licenses/, Accessed May 2016.

[9] Elvira Mitraka, Andra Waagmeester, Sebastian Burgstaller-Muehlbacher, Lynn M. Schriml, Andrew I. Su, and Benjamin M. Good. Wikidata: A platform for data integration and dissemination for the life sciences and beyond. In *Proceedings of the 8th International Conference on Semantic Web Applications and Tools for Life Sciences (SWAT4LS)*, 2015.

[10] Julie McMurry, Juha Muilu, Michel Dumontier, Henning Hermjakob, Nathalie Conte, Philipp Gormanns, Murat Sariyar, Janna Hastings, Alejandra Gonzalez-Beltran, Niklas Blomberg, Chris Morris, Jean-Karim He´riche´, Melissa A Haendel, Rafael C Jimenez, Tony Burdett, Philippe Rocca-Serra, Nicolas Le Nove`re, Nick Juty, Katherine Wolstencroft, Simon Jupp, Wolfgang Mu¨ller, Donal K Fellows, Maria J Martin, Neil Swainston, Helen Parkinson, Carole Goble, Johanna R McEntyre, Camille Laibe, Jacky L Snoep, Nicole Washington, Susanna-Assunta Sansone, Natalie J Stanford, Jon C Ison, Alan R Williams, Christopher J Mungall, and James Malone. 10 Simple rules for design, provision, and reuse of persistent identifiers for life science data. Available from http:// zenodo.org/record/18003. May 2015.

[11] GO Curator Guide. Available from http://wiki.geneontology. org/index.php/Curator_Guide:_Obsoletion. Accessed May 2016.

[12] Ontobee webserver. Available from http://www.ontobee.org. Accessed May 2016.

[13] Creating an ontology project, an update. Available from https://douroucouli.wordpress.com/2015/12/16/creating-an-ontology-project-an-update/. Accessed May 2016.

[14] Ontology metadata common set. Available from http://information-artifact-ontology. googlecode.com/svn/releases/2015-02-23/ ontology-metadata.owl. Accessed May 2016.

[15] W3C. Simple Knowledge Organization System (SKOS). Available from http://www.w3.org/TR/2009/ REC-skos-reference-20090818/, Accessed May 2016.

[16] Dublin Core Metadata Initiative. Dublin Core Metadata Element Set. Available from http://dublincore.org/documents/ dces/, Accessed May 2016.

[17] Olga Vrousgou, Tony Burdett, Simon Jupp, and Helen Parkinson. Biomedical ontology evolution in the embl-ebi ontology lookup service. In *Proceedings of the Workshops of the EDBT/ICDT 2016 Joint Conference (EDBT/ICDT 2016)*, 2016.

[18] OBO Foundry principles - overview. Available from http://obofoundry.org/principles/fp-000-summary.html. Accessed May 2016.

[19] A Bandrowski, R Brinkman, M Brochhausen, MH Brush, B Bug, MC Chibucos, K Clancy, M Courtot, D Derom, M Dumontier, et al. The ontology for biomedical investigations. *PloS one*, 11(4):e0154556, 2016.

[20] Uberon Design patterns. Available from https://github. com/obophenotype/uberon/wiki/Manual#design-patterns. Accessed May 2016.

[21] Bard, Jonathan, Seung Y Rhee, and Michael Ashburner. "An Ontology for Cell Types." *Genome Biology* 6.2 (2005): R21. *PMC*. Web. 30 June 2016.

[22] Heiko Dietze, Tanya Z Berardini, Rebecca E Foulger, David P Hill, Jane Lomax, David Osumi-Sutherland, Paola Roncaglia, and Christopher J Mungall. Termgenie– a web-application for pattern-based ontology class generation. *Journal of biomedical semantics*, 5(1):1, 2014.

[23] Phillip Lord. The semantic web takes wing: Programming ontologies with tawny-owl. *arXiv preprint arXiv:1303.0213*, 2013.

[24] Mikel Egana, Alan Rector, Robert Stevens, and Erick Antezana. Applying ontology design patterns in bio-ontologies. In *Knowledge Engineering: Practice and Patterns*, pages 7–16. Springer, 2008.

[25] Simon Jupp, Tony Burdett, Danielle Welter, Sirarat Sarntivijai, Helen Parkinson, and James Malone. Webulous and the webulous google add-on-a web service and application for ontology building from templates. Journal of Biomedical Semantics, 7(1):1, 2016.

[26] Wikipedia - Upper Ontology. Available from https://en.wikipedia.org/wiki/Upper_ontology#Available_ontologies. Accessed May 2016.

[27] P. Grenon, B. Smith, and L. Goldberg. Biodynamic ontology: applying bfo in the biomedical domain. *Studies* in health technology and informatics, 102:20–38, 2004.

[28] The Goldilocks principle. Available from https://en.wikipedia.org/wiki/Goldilocks_principle, Accessed June 2016.

[29] Phillip Lord. Ontogenesis. Available from http://ontogenesis.knowledgeblog.org/, Accessed May 2016.