Evaluations for Choosing Terms and Ontologies (Real Live Version)

Background

This document is intended to encourage community awareness of term and ontology selection approaches. Whether it becomes a white paper, on-line recommendations, and/or publication is TBD.

To contribute, put your name in the author list following the pattern. Then use Suggesting *if you propose significant modifications to what's there or want it explicitly reviewed. Use* Edit *mode for minor changes.*

Bitly Link: https://bit.ly/evaluations-choosing-terms-ontologies

Authors

John Graybeal, Consultant (VSSIG Co-Chair), jbgraybeal@sonic.net (0000-0001-6875-5360)
Asiyah Lin, Axle Research Technology & NIH, asiyah.lin@nih.gov (0000-0003-2620-0345)
Anna Maria Masci, MD Anderson Cancer Center University of Texas, amasci@mdanderson.org (0000-0003-1940-6740)
Juliane Schneider, PNNL, juliane.schneider@pnnl.gov (0000-0002-7664-3331)
Hande Küçük McGinty, Kansas State University, hande@ksu.edu (0000-0002-9025-5538)
Eric G Stephan, PNNL, eric.stephan@pnnl.gov (0000-0002-8155-6806)

Goals

Collect a reasonably comprehensive list of metrics and indicators of the value of a term or ontology, including characterizing the fitness for specific purposes of those metrics and indicators. Discuss possible sources of the needed data for those metrics. Reflect how the metrics can be adjusted (either in weighting, or in content) for particular use cases. Provide examples of how the metrics might be, or are, collected, presented, and discovered.

Revisions

Date	Person	Description
2024.06.17	John G	Added ontology reuse implications. Copied all OBO Foundry principles into 'header advice' of appropriate

		sections, along with adaptations for our general case.
2024.06.03	John G	Re-migrated latest content to new doc, disabled old doc
2024.05.20	Asiyah L	Added the section of repurpose/reform/recycle ontology when see fits in use cases.
2024.04.22	Juliane S	Added some draft ideas under Matching Criteria
2024.04.08 2024.04.10	Eric S	Added Governance Criteria content, Reuse of Profiles, Reuse of Single Terms, Reuse of Ontology as a whole
2024.03.26	John G	Added Internationalization Criteria, outlined Analytics metrics, and create an external table of metrics
2024.02.26	Anna Maria/John G	Added discussion of Popularity Criteria
2024.02.13	John G	Added more use cases, and this revision table

Table of Contents

Access through Bitly link
https://bit.ly/evaluations-choosing-terms-ontologies

Introduction

One of the biggest challenges for new users of semantics, and for many others with more experience, is finding good terms and good ontologies. By definition, a good ontology will have good terms in it—but they may not be the good terms you need, even if the ontology is about your topic. And there might be some excellent terms hidden in obscure ontologies. Even if you can find all the options, how can you decide which ones are best for your needs?

It turns out that humans, especially scientific and engineering humans, need to use a lot of different terms for all sorts of activities. So finding a "best match" may never be simple, since as coverage of a given topic increases, so do the number of artifacts to choose from, and the complexity of each artifact.

Intuitively people think this must be a straightforward problem to solve, that there should be an algorithm that can do the choosing for them. Although there are many algorithms to match a term, most are based simply on label matching, and perhaps indirectly some other factors. There are no known algorithms that reflect any significant aspects of the complexity shown here.

However, all is not lost. With the basic algorithms that are implemented in a few tools, and using those tools and other guidelines, with human judgment you can learn to make relatively good judgments about terms and ontologies relatively quickly.

In this white paper we illustrate the challenge, and offer a broad proposal for algorithms that could effectively help people quickly find the best terms and ontologies for their purposes.

The Alternative to Reuse: Developing Your Own Ontology

Most people who develop their own ontology have at least looked at existing ontologies to see if their needs might be met. While ideally ontology development would rarely be necessary, realistically existing semantic artifacts are often insufficient to meet complex or scientifically novel requirements. Nonetheless, our white paper can be a useful contribution to this use case in several ways.

First, the evaluation criteria for choosing ontologies can also make very clear why existing ontologies are not sufficient for the new use case. Technical, social, and other factors are often cited for not using existing ontologies, but precise metrics and principles can more thoroughly inform and justify the decision.

Second, the new ontology will typically have to be related to the existing semantic artifacts addressing the same domain(s). Our treatment will inform the choices about ontologies to map, and term mappings to create.

Third, by laying out principles for selecting ontologies, we hope to motivate the managers of ontologies to satisfy more of those principles in their own products, in order to see their products used and cited more often. Greater emphasis on best practices in standardization, structure, and formality may bring more ontologies up to the level of “good enough” for future users, saving considerable effort and unnecessary diversification of semantic products.

Finally, by opening up more opportunities and techniques to find, evaluate, and repurpose not just whole ontologies, but ontology components and terms, we hope to enable a larger and more diverse market for ontology adoption.

Catalog of Evaluation Criteria

This section describes criteria the authors have identified that could help you understand whether a term or ontology is a good fit for your needs.

Categories of Evaluation Criteria

Each of these categories has multiple specific criteria (many objective, some subjective or harder to measure) that could be used to evaluate it.

- Popularity (including reuse, visits, selections from a search list, other voting techniques)
- Reuse of ontologies / terms (in either case, reuse considers the ontology as a whole, ontology fragments, design patterns, profiles, and individual terms)
- Governance (methods and controls for modification and change tracking; trust indicators; change frequency and recency)
- Community (as in, domain) relevance (adoption, standardization, concept applicability)
- Internationalization (how widely used, 'level' of governing body, international re-use, use of multiple languages in text annotations)
- Matching (precision, and if multiple terms, match frequency and specialization)
- Quality (by metrics, by independent assessment(s))
- Analytics (ontology structure)
- Best Practices Adoption (documented recommendations on ontology/term creation; includes FAIRness recommendations)

The complete list of criteria is available in spreadsheet format at the Table 1 spreadsheet Evaluation Criteria Table.

Popularity Criteria

Popularity (including reuse, views, selections from a search list, other voting techniques)

In the context of ontology evaluation, the criterion of popularity bears significant weight, primarily due to its influence on promoting a shared terminology within the scientific community. However, this reliance on popularity harbors a dual-edged nature. The crux of the issue lies in the perpetuation of a potentially detrimental cycle, where ontologies of inferior quality gain traction and widespread use simply because of their popularity. This scenario not only undermines the integrity of shared knowledge but also impedes the adoption of more accurate and reliable ontologies.

A pertinent illustration of this problem can be observed through the example of searching for "macrophage" in the NCBO BioPortal. It reveals that the Cell Ontology, despite being a domain-specific ontology tailored for cell types, does not appear as the primary result. Such a situation poses a risk for users who are either unacquainted with the intricacies of ontology selection or those momentarily inattentive, leading them to opt for the first, potentially less suitable option presented. This inadvertently fuels the cycle of reusing suboptimal ontologies due to their apparent popularity. Addressing this issue is crucial, not only to mitigate the propagation of inaccuracies but also to elevate awareness among users regarding the critical need for discerning evaluation of ontologies before their application.

Popularity metric: Views

(many sources of 'View' information)

Popularity metric: Reuse

This is covered in detail below. The idea here is that the amount of reuse can also be a factor in popularity calculations. Since this amounts to double-counting, and in implementation could reduce the user's control over the inclusion of reuse in the final evaluations, we will not consider reuse further in the popularity discussion.

Popularity metric: User Selection

(note the difficulty of measuring, since generally this happens in external tools)

Popularity metric: Votes and Reviews

(happens in higher levels, e.g., ontologies and maybe branches; but not so often at level of terms)

Popularity metric: Other attention metrics

(for example, number of changes over time; number of requests for changes; web searches that match the label; occurrences of identifier 'in the wild')

Reuse Criteria

OBO P9) Documented Plurality of Users - The ontology developers should document that the ontology is used by multiple independent people or organizations.

Many semantic applications benefit from re-using semantic assets that already exist. In this section we review both common and subtle scenarios for this evaluation.

Perhaps the most important indicator of whether an ontology or a term is suitable for re-use is how many times the ontology or term has already been re-used. At the simplest level, some metric may exist of the popularity of ontology re-use, for example page visits to the ontology's summary page or a count of its downloads. Terms are less easily monitored, but some tools may provide an indication of their activity or adoption as well. For example, when BioPortal provides a list of terms satisfying search criteria, it prioritizes and groups terms that have been repeated in other ontologies, and also favors terms in popular ontologies.

In some applications, the only unit of re-use is an entire ontology, but other applications understand ontologies and can use that understanding to select parts of the ontology, including: individual terms (by themselves, or including annotations about them); coherent fragments of an ontology, for example a term and all the terms below it in a *subClass*, *partOf*, or *broader/narrower* relationship; terms and annotations that satisfy a particular design pattern, such as a query pattern; or a subset of terms that have been collected in a separate ontology, known as an *ontology profile* (also know as an Ontology View).

In cases where an application only supports re-using an entire ontology—for example, in the semantic import mechanism itself, owl:imports—users will create a collection of statements that meets their need to select parts of an ontology. The application can then treat this collection as a complete ontology and work with it directly.

Here we consider how re-use of ontologies and their selected parts can be measured and used to determine the value of a particular ontology or partial ontology.

Reuse of ontology as a whole

Standards adoption is a major driver for ontology reuse, by adopting an existing ontology, implementors leverage the work already vetted by domain experts. Reuse of an existing ontology can take different forms including: off-the-shelf (OTS), federation, extension, translation, and evolution. The off-the-shelf approach may be most common for practitioners to reuse an ontology within the original domain for which it was intended. Federation is reuse shared between two distinct domains, it is typically composed of two or more loosely coupled ontologies (*consider ontology taxonomies as well*). Reusing an upper or enterprise ontology with extended classes supports rapid customization for enterprises via extensibility, and *owl:import* provides loosely coupled modularity supporting federation. When ontologies are

managed through community support (see Governance Criteria) the practice of versioning ontologies due to evolution should be considered as another type of reuse where communities build upon core concepts.

Reuse of ontology fragments (anything bigger than a term and its annotations)

Interrelated terms commonly used together to implicitly convey a concept are considered ontology fragments. Examples include: address (street, city, territory, country), contact information (name, email address, phone number(s)). Collectively ontology fragments can be used to represent a specific domain's concepts, relationships, and constraints.

> Repurpose/reform/recycle ontology fragments when see fits in use cases
>
> Based on the actual use case needs, sometimes an ontology is being chosen as the primary base ontology to modify, including change the hierarchies, modifying the definitions, and even term labels. As the ontology often have an open source license, such modifications are allowed. In this case, to keep the provenance of the ontology use is important.

Reuse of ontology design patterns

Reuse of ontology profiles (?)

Ontology profiles also referred to as application/data profiles are collections of ontology terms derived from existing ontologies used to specify data exchanges, data processing, and databases. While profiles reuse existing ontology terms, the greatest benefit is their ability to provide applications and systems with verifiable and exact ontology content. While the DX-Profile vocabulary does exist, profiles can be generated by creating a smaller ontology relying on OWL:Import, as well as using the RDF Shape Constraint Language (SHACL). Profiles are modular making it possible for one profile to be composed of several components. Profiles can be published by standards communities in OWL, but are also commonly published as technical documentation and behind paywalls making their reuse metrics a bit more elusive to collect by external communities.

Reuse of individual terms

Ontological terms are designed to have specific meanings and that can be applied to different domains and contexts. Reusing ontology terms provides consistency as well as interoperability between different applications. Explicit term reuse involves the use of a term based on its original unique identifier or to support terminology refinement, creating a child term inherited from the original term's unique identifier, or motivated via native language translation creating an equivalent term with a matching definition and structure pattern of the original term. Implicit reuse… (should we mention implicit anti-patterns and the potential for false positives or improper duplication of efforts?)

Governance Criteria

methods and controls for modification and change tracking; trust indicators; change frequency and recency; openness of processes and participation
OBO P1) Open - The ontology MUST be openly available to be used by all without any constraint other than (a) its origin must be acknowledged and (b) it is not to be altered and subsequently redistributed in altered form under the original name or with the same identifiers.
OBO P10) Commitment To Collaboration - ~~OBO Foundry~~ ontology development, in common with many other standards-oriented scientific activities, should be carried out in a collaborative fashion.
OBO P13) Notification of Changes - Ontologies SHOULD announce major changes to relevant stakeholders and collaborators ahead of release.
OBO P16) Maintenance - The ontology needs to reflect changes in scientific consensus to remain accurate over time.
OBO P20) Responsiveness - Ontology developers MUST offer channels for community participation and SHOULD be responsive to requests.

Mature ontology governance indicators may vary depending on the industry and domain. Some common characteristics include providing transparency of the governance body structure, and longevity of participating organizations and activities of the membership that can come in the form of active task force meetings that support proposals and issue tracking. Mature standards communities are built upon documented consensus-based processes for developing and updating the ontology based on a lifecycle approach including approval processes that are tracked supporting either versioned periodic releases of an updated ontology, or continuous integration strategy where elements of the ontology are updated and versioned. Other criteria include global recognition where the ontology is referenced or adopted by standards organizations or where the ontology governing body is actively collaborating with other standards organizations to support interoperability and harmonization.

Community Relevance Criteria

(As in domain communities) adoption, standardization (defined by community's needs),enforcement (required by community edict), concept applicability, cultural considerations, communications (supported by community)
P11) Locus of Authority - There should be a person who is responsible for communications between the community and the ontology developers,~~ for communicating with the Foundry on all Foundry related matters,~~ for mediating discussions involving maintenance in the light of scientific advance, and for ensuring that all user feedback is addressed.

Internationalization Criteria

The criteria in this section reflect content support for internationalization, and social aspects of international adoption. In short, these criteria consider the artifact's community relevance for the international community.

International content (multi-language annotations, IRIs)

Concrete metrics of the level of internationalization consider whether the ontology has been designed for use internationally, particularly in multi-lingual settings. Several specific criteria can be evaluated.

Multi-language annotations

Do the ontology's text annotations—particularly labels and definitions—include content in multiple languages, annotated with the language of the text string? For example we can express someone's title in two languages as

```
<https://ex.org/person/John-Doe-38765> foaf:title "Director"@en, "Directeur"@fr .
```

or even include a regional annotation like

```
<https://ex.org/menuitem/23> rdfs:label "french fries"@en-US, "chips"@en-GB .
```

If the content is multi-lingual, we can expect to see most of the string literals annotated with language strings, and possibly regional strings. The level of internationalization can be considered proportional to the number of different languages and their frequency of use—if there are many languages represented, and they are all present for all of the literal strings, the document is highly internationalized.

Semantic IRIs [anti-pattern]

An identifier in our ontology that uses semantically meaningful strings—particularly for the final fragment—is expressing a preference for the language represented in the IRI, at the expense of other languages. While this is unavoidable in the domain name, it is more avoidable in the rest of the namespace, and very avoidable in the identifier fragment. While this choice does not break internationalization, it disfavors non-ASCII languages, for example when trying to represent a term that contains a non-ASCII Unicode character, either the Unicode character must be dropped, converted to the best-match ASCII character, or represented using a Unicode escape sequence. Any of these solutions creates confusion for representing and working with most languages, which can be easily avoided by not using semantics in the IRI.

International governance

Ontologies may be governed by individuals, organizations, or communities. A strong indication of their internationalization level is the geographic scope of the governance body and process. In the case of an individual, the governance body can not be measured, and more weight should be given to the process.

A governance body that is composed of more representation from more countries can be considered more international, especially if those countries are not from a small region. A governance body from a single country, or a small group of countries, can not be counted on to bring an international perspective to the process. Participation in communities and governmental bodies may evolve over time, but the probability over time is that more international bodies will produce more international ontologies.

To evaluate the geographic scope of the ontology governance process requires that the process be explicitly defined, and that the definition and execution be open to inspection. An international process will include methods for contributors from any country to submit their contributions and have them considered, ideally with input from an internationally composed set of reviewers.

International visibility and re-use

Ontologies that are used across broader geographical and political boundaries have larger universality, and therefore larger value for knowledge specification. These traits can be measured in several formal and informal ways, not always through direct automated metrics.

A feasible international re-use metric is the prevalence of an ontology in papers written in different languages, or by authors representing institutions from different locations. Citation of ontologies (or papers about them) in papers that represent more languages and locations indicate broader international adoption. Conversely, ontologies addressing overlapping topics that are cited in more specific languages or locations are almost certainly not internationally adopted.

Where national standards bodies or localized domain communities are involved, the adoption of ontologies specified by the governance organization signals the level of international re-use. For example, EU bodies may specify one set of terminologies for a particular domain, while US organizations use a different set. These differences are not explicitly tracked, but are identifiable by experienced practitioners in the domain.

Relative international visibility is another indicator that can be measured, but is not necessarily published in existing statistics that are widely known. For example, in many cases ontologies are translated into another language which is published independently of the original. The secondary ontology may be presented in a language-specific ontology repository, and brought forward by language-specific search engines. Such a secondary ontology will naturally be used predominantly in the countries that speak that language, while the initial ontology may be used little or not at all in the countries where the translated ontology is more visible, thereby reinforcing the visibility differences. Visibility might be measured by discovery characteristics: how highly placed is the ontology and its publications in search engines, search tools, and ontology repositories?

Matching Criteria

For ontology: the right topic. Within an ontology: Precision, and if multiple terms, match frequency and specialization;

With respect to individual terms: level of context considered in matching process; definitional checks;

Definitions in an ontology should be supported by domain literature, and the source cited in the term properties. If not, the context is unclear. For instance, the term ‘antenna’ could be an anatomical part of a butterfly, or it could be a sensor on a piece of equipment.

The level of granularity in an ontology should be consistent across the branches or knowledge graph.

Rich synonomy should be incorporated into an ontology for precision in term matching

Analytics

Information about ontology structure

Ontology analytics can provide many metrics that summarize an ontology, and in many cases those statistics can indicate whether the ontology is suitable for a particular purpose. Unlike some of the other criteria categories, analytics results usually can not be interpreted independently of the use case.

To take a common example, the flatness of an ontology suggests its complexity, with very flat ontologies (1 or a few parent terms and a very large number of children under each parent, with no depth beyond that) being conceptually very simple. For most purposes an extremely flat ontology is not powerful, and therefore not useful. But if all you need is a long list of terms in a particular domain, for example to search for string matches, an all-encompassing flat ontology could be the ideal solution.

In this section we attempt to gather most of the well-understood ontology analytic measures. We roughly group them from simplest to most complex, keep the descriptions of each as simple as possible, and offer some notion of how to interpret the measure.

Counting Metrics

Number of entities, collectively or by type [class, property, annotation, individual, …]

Number of annotations by type [label, definition, broader, narrower, isA, partOf, synonym/mappings…]

Number of namespaces

Number of metadata attributes

Number of included ontologies

Number of foreign entities used

Simple Calculations

Percentage of terms with labels and/or definitions

Maximum depth of hierarchy (determined by isA, partOf, or broader/narrower relations)

Maximum number of children

Flatness (depth vs breadth)

Percentage of foreign entities used

More Complex Calculations

Need help here…

Ontology shape (?)

Connectedness (? - more of a graph or a tree)

Patterns and Other Complex Assessments

Need help here…

Credibility of foreign terms

Quality

by metrics, by independent assessment(s). Note OBO Foundry criteria here
OBO P2) Common Format - The ontology is made available in a common formal language in an accepted concrete syntax.

OBO P5) Scope - The scope of an ontology is the extent of the domain or subject matter it intends to cover. The ontology must have a clearly specified scope and content that adheres to that scope.
OBO P6) Textual Definitions - The ontology has textual definitions for the majority of its classes and for top level terms in particular.

Best Practices Adoption Criteria

documented recommendations on ontology/term creation; includes FAIRness recommendations, Versioning style of ontology and terms
OBO P3) URI/Identifier Space - Each ontology MUST have a unique ***persistent resolvable*** *IRI ~~in the form of an OBO Foundry permanent URL (PURL).~~*
OBO P4) Versioning - The ontology provider has documented procedures for versioning the ontology, and different versions of ontology are marked, stored, and officially released.
OBO P7) Relations - Relations should be reused from the Relations Ontology (RO) ***and/or other well-managed community ontology whenever possible.***
OBO P8) Documentation - The owners of the ontology should strive to provide as much documentation as possible.
OBO P12) Naming Conventions - The names (primary labels) for elements (classes, properties, etc.) in an ontology must be intelligible to ~~scientists~~ ***domain users*** *and amenable to natural language processing. Primary labels should be unique among ~~OBO Library~~ ontologies* ***in use****.*

Table 1. Criteria by Category

Existing Evaluation Systems

Technologies/Models

Implemented technologies (or proposed approaches) for evaluating the quality, usability, and/or fitness of semantic assets

Term Evaluations

- syntactic match
- (inherited) value of enclosing ontology
- identifier match
- desired domain (can come from enclosing ontology)
- Machine Learning techniques
- Large Language Model techniques
- Existing (documented) matches or mappings, through annotations, mappings, etc.)

Ontology Evaluations

- BioPortal Recommender (version 2)
    - match frequency
    - specialization of ontology
    - popularity (as measured locally)
- See OBO evaluation criteria at http://dashboard.obofoundry.org/
- Level of reuse
- Ontology reviews (as formerly used in BioPortal)
- Metadata adoption

Proposed Evaluation Facets

Not clear what we should include here. Possibly our recommendations of best approaches, but we may not be able to even scope this until we have finished more of the above sections.

Term Evaluations

- Closeness of match
    - preferred label
    - synonyms, alternate labels
    - description (text describing the meaning of the term for human readers)
    - relations and annotations

Ontology Evaluations

- Basic ontology structure
    - Is it included as a base ontology in other ontologies?
    - Is it a stand-alone ontology or does it depend on other ontologies?
- What is its topic relative to the desired topic? ("I want an ontology about X")
- Architectural considerations
    - Patterns and frameworks used
        - Hierarchical relations used
        - Multiple inheritance
        - Strict .. (umm, top-down tree structure?)
        - Rule-based classes (technical term, please?)
    - Metrics indicators
        - Max depth of hierarchy
        - Average # child nodes
        - Entity types and relative frequencies (classes, properties, individuals, annotations)
        - (Review existing metrics in BioPortal, other repos)

Term-Ontology Interactions

- Adding terms to an existing ontology
- Composition of terms to satisfy a need

Use Cases to Discuss

When evaluating the criteria, it is important to appreciate the exact goal of selection activity. This section lists a number of precise goals that might lead to searching for a term or ontology.

Searching for a term to…

Learn what the term means

Explore or narrow the context being searched

Document a specific aspect of an object (e.g., a data set)

Find the authoritative definition of that term or IRI

Help formulate a data shape or semantic profile

Searching for a set of terms

Name attributes to describe an object

Characterize current state of research

Searching for an ontology

Annotate free text with controlled terms from a particular domain

Specify terms to use in describing data or processes

Learn more about a particular domain and/or its terminology

Find a good home for a term that doesn't have a good home

Searching for a set of ontologies

Search for a set of ontologies (to annotate free text with controlled terms from multiple domains)

Search for a set of ontologies (to characterize the domains represented by a free text collection)

Evaluation Guidance

How should these different components be weighted and prioritized? Are there some overarching approaches that are likely to be effective generally, or in well-defined scenarios?

Today: By Yourself With Current Tools

Future: In A Better System

Far Future: In An Ideal System

Future Tasks

What else can be included to advance this analysis?

Add Section: Choosing an existing term or ontology vs creating your own

Create: Proof of Concept

Create: Interfaces for running metrics

Resources

Many references related to ontology and term creation and selection have been collected at Vocabulary References focusing on SKOS. That document attempts to bring to the top the most useful advice and tools for the early-stage user of semantic resources, and also references more sophisticated content where it adds practical value.

Ideally this section can perform the same function for choosing terms and ontologies, quickly exposing users to more advanced techniques, tools, and lessons.