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.

This project is managed through the Research Data Alliance Vocabulary and Semantic Services Interest Group. 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. If you are one of the editing team, you can use* Edit *mode for minor changes.*

To join the team, contact John Graybeal at jbgraybeal@sonic.net, and/or visit the ontochoice.github.io web site to learn how to volunteer. You may also want to visit the VSSIG homepage to learn how to join the VSSIG slack workspace, then join the #tg-term-selection channel.

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

QR Code to Document:

Posters

- Poster submitted to ICBO and FOIS
- ESIP Poster

Authors

John Graybeal, Consultant (VSSIG Co-Chair), jbgraybeal@sonic.net (0000-0001-6875-5360)
Asiyah Lin, OntoData Research and Solutions, LLC., ontology.world@gmail.com (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)
Eric G Stephan, PNNL, eric.stephan@pnnl.gov (0000-0002-8155-6806)
Wolmar Nyberg Åkerström, wolmar.n.akerstrom@uu.se (0000-0002-3890-6620)
Barbara Magagna, barbara@gofair.foundation (0000-0003-2195-3997)
Muhammad Tuan Amith, muamith@utmb.edu (0000-0003-4333-1857)
Maria Shatz, NIEHS/NIH, maria.shatz@nih.gov (0000-0002-6981-5174)
Hande Küçük McGinty, Kansas State University, hande@ksu.edu (0000-0002-9025-5538)

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.

Organization of Document

This preface material includes an introduction for new readers, a list of authors, and the goals of the document, followed by the table of contents.

After that the main document begins with the Introduction on page 1.

In the main document and Table of Contents, colored circles before the section headers indicate the maturity of that section. The status key for those is as follows:
🟢 : Section is first-draft complete. (Additions/improvements still welcome and likely needed.)
🔰 : Section has useful text and guidance but needs work.
🔶 : Section has outline topics/example only.
⭕ : Section has little useful content.
🔲: Need for section is unclear; may be deleted or moved.

Key to interpreting text

Black regular or bold text is intended as (rough) final copy.

Black italicized text has changed since the last significant revision. The italics will be removed once a significant review has occurred by at least one person.
Gray italicized text is preface material to describe a particular section or subsection. It is removed once it has been addressed by black regular text.

Revisions

Date	Person	Description
2025.10.06	John Graybeal	Updated sections D,E,F to green (after suitable modifications), improvements in several other sections, (also created list of tasks needed to reach 'Zenodo status')
2025.07.13	John Graybeal	Added Tuan to author list
2025.04.21	John Graybeal, Anna Maria Masci	Updated impact of criteria section to get/link to Google sheet; first take on Guidance section
2025.04.07	John Graybeal	Incorporated edits/comments from Hande, Maria, Anna Maria, Barbara. Added significantly to Community Approval Criteria, Popularity Criteria, Quality Assessed Criteria. Added section in Use Cases to address impact of criteria in addressing specific use cases.
2025.02.07	John Graybeal	Added considerably to (A) Right Topics and Terms (Matching) Criteria >> Ontology Topic(s) section
2025.01.27	John Graybeal	Added key to interpreting text and formatting to follow it. (A) Added a lot about 'Right Topic(s)' and subsection re synonyms (G) Twiddled best practices to discuss granularity (J) Added content about responsiveness
2025.01.21	John Graybeal	Updated first part of internationalization section, inspired by comments from Barbara M and Nico M
2025.01.16	John Graybeal	Changes from RDA VSSIG session: (F) Importance of context in determining reuse (H) Added quality evaluations, integration with other ontologies (J) Added operational responsiveness criteria (new) Application of LLMs
2024.11.10	John G/Hande	Inserted reference/slide from outside presentation
2024.10.21	John G	Updated Typical Evaluation Sequence section, and

		Categories C and D
2024.09.16	John G	Added emoji status indicators. Also updated a few sections to be more contributor-ready.
2024.08.31	John G	Reorganized/reordered Categories to align with posters. Added information from ICBO poster to Relevance group.
2024.08.26	John G	Minor editorial changes…
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

Background 1
- Authors 2
- Goals 2
- Organization of Document 2
- Key to interpreting text 2
- Revisions 3
- Table of Contents 5

Introduction 1
- 🟢 Challenge and Approach 1
- 🟢 The Alternative to Reuse: Developing Your Own Ontology 2

Categories of Evaluation Criteria 3
- 🟢 Evaluation Categories List 3
- 🟢 Typical Evaluation Priority 3

Category Descriptions and Detailed Criteria: Relevance Group 5
- 🔰 (A) Right Topics and Terms (Matching) Criteria 5
- 🟢 (B) Required Ontology Structure Criteria 10
- 🟢 (C) Community Enforced Selection Criteria 10

Category Descriptions and Detailed Criteria: Popularity and Reuse Group 11
- 🟢 (D) Community Approval Criteria 11
- 🟢 (E) Popularity Criteria 13
- 🔰 (F) Reuse Criteria 16

Category Descriptions and Detailed Criteria: Best Practices and Analytics Group 19
- 🔶 (G) Best Practices Adoption Criteria 19
- 🔰 (H) Other Quality Assessment Criteria 21
- ⭕ (I) Value-Neutral Analytic Criteria 23

Category Descriptions and Detailed Criteria: Governance and Internationalization Group 25
- 🔰 (J) Governance Criteria 25
- 🟢 (K) Internationalization Criteria 26

Existing Evaluation Systems, Technologies, and Models 29
- 🔶 Existing Tools 29
- 🔶 Implemented Technologies 30
- 🔶 Published Approaches (Not Yet Implemented) 30

Applicability of Large Language Models (LLMs) 31
- ⭕ Ability of LLMs to Process and Explain Existing Terms and Ontologies 31
- ⭕ Use of LLMs to Recommend Terms and Ontologies 31
- ⭕ Pre-conditioning LLMs With Existing Schema Before Recommending 31

⭕ Generating Ontology Content Directly with LLMs 31
Recommended Evaluation Facets 31
⭕ Term Evaluations 32
⭕ Ontology Evaluations 32
⭕ Term-Ontology Interactions 32
Use Cases and Their Impact 32
🔶 Searching for a term to… 33
🔶 Searching for a set of terms 33
🔶 Searching for an ontology 34
🔶 Searching for a set of ontologies 34
🔶 Sources for additional use cases 34
🔶 Use Case Impact Heat Map 34
Evaluation Guidance 35
🔰 Discovery Paths 35
🔰 Selection Now: By Yourself With Current Tools 36
🔶 Selection Future: In A Better System 37
🔶 Selection Far Future: In An Ideal System 38
Future Tasks 38
🔲 Add Section: Choosing an existing term or ontology vs creating your own 38
⭕ Create: Proof of Concept 38
⭕ Create: Interfaces for running metrics 38
🔶 Resources and Bibliography 38
Resources 39
Bibliography 39

Introduction

🟢 Challenge and Approach

One of the biggest challenges for new users of semantics, and for many others with more experience, is finding good terms and good ontologies. (In this paper we use ‘ontology’ to mean any semantic resources described using W3C semantic standards.) 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?

This task is especially challenging for a researcher who has not engaged with the ontology community for a long time (or ever), or has multiple needs for choosing and reusing ontologies. For example, the same ontology may be needed for both creating a proprietary knowledge graph, and performing a natural language processing (NLP) task.

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 attention to a given topic increases, so does the number of artifacts to choose from, as well as 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 the complexity described 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, and some ontology creation workflows actively encourage reusing existing ontologies as part of their workflow (KNARM[1], OLIVE[2], SPIRES[3]).

And while ideally ontology development would rarely be necessary, realistically existing ontologies are often insufficient to meet complex or scientifically novel requirements. Therefore, our white paper can be a useful contribution to ontology development in several ways.

First, the evaluation criteria for choosing ontologies also demonstrate 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 which term mappings to create.

Third, by laying out principles for selecting ontologies, ontology developers may try 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 can bring more ontologies up to the level of “good enough” for future users, saving considerable effort and unnecessary diversification of semantic products.

Fourth, these principles may help tool and community creators to address the needs of ontology creators and users to find and re-use existing semantic artifacts. By building in knowledge of good semantic practices, these tools can reduce the effort and increase the quality 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.

---

[1] McGinty, Hande Küçük. Knowledge acquisition and representation methodology (knarm) and its applications. Diss. University of Miami, 2018.
[2] upcoming paper at Applied Ontology Journal
[3] Caufield, J. Harry, et al. "Structured prompt interrogation and recursive extraction of semantics (SPIRES): A method for populating knowledge bases using zero-shot learning." Bioinformatics 40.3 (2024): btae104.

Categories of Evaluation Criteria

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

However criteria may be weighted, in most scenarios the user can determine the suitability of a term or ontology early in the evaluation process, given the limited number of terms and ontologies available for each domain.

🟢 Evaluation Categories List

Each of these categories has multiple specific criteria that could be used to evaluate the ontology or term. Many of the criteria are objective, some criteria are relatively hard to measure, and some criteria are subjective. The categories are listed below, with criteria examples in parentheses.

A. Right Topics and Terms (matching precision, and if multiple terms, match frequency and specialization)
B. Required Ontology Structure (usable size and patterns, compatible ontology structure)
C. Community-Enforced Selection (a selection the relevant community requires to be used)
D. Community-Approved (adoption, standardization, concept applicability)
E. Popularity (including reuse, visits, selections from a search list, other voting techniques)
F. Reuse (in either case, reuse considers the ontology as a whole, ontology fragments, design patterns, profiles, and individual terms)
G. Best Practices Adoption (documented recommendations on ontology/term creation; includes FAIRness recommendations)
H. Measured Characteristics (ontology structure, shape, and other measured characteristics)
I. Quality Assessed (by independent human assessment)
J. Governance (methods and controls for modification and change tracking; trust indicators; change frequency and recency; broad participation in governance)
K. Internationalization ('level' of governing body, international re-use, use of multiple languages in text annotations)

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

🟢 Typical Evaluation Priority

To illustrate the typical scenario in which a user is determining whether a term or ontology is suitable, this section groups the major categories and presents those groups in a potential

sequential order of their application. The organization begins with fundamental weed-out criteria, and proceeds to progressively less important criteria for typical use cases.

The outline below and the associated Figure 1 embeds a typical sequence of evaluation categories and user priorities. The top-level items are considered category Groups, and the second level items are the Categories themselves.

The first Group, Relevance, includes criteria that narrow the search to those terms or ontologies which have at least some potential to satisfy the user's needs. The remaining three Groups refine the prioritization when multiple terms or ontologies remain under consideration.

Because the last 3 Groups may be weighted differently by different user communities, or for different user scenarios, evaluation tools should always allow weighting of the Groups, Categories, and/or individual criteria. With such weighting available in an evaluation tool, it becomes possible for anyone to emphasize the evaluations they consider most important, and ignore criteria they deem unimportant.

The following outline shows the Category Groups and Categories. The parenthetical letters after each second-level category refer to the Category list in the Evaluation Categories List section above. The same organization is used to organize and discuss the categories and criteria, in the four sections that follow.

- Relevance (These are simple weed-out criteria)
    - Right Topics and Terms (A)
    - Required Ontology Structure (needed patterns, size, and structure) (B)
    - Community Enforced Selection (C)
- Popularity and Reuse
    - Community Approval (D)
    - Popularity and Reuse (E,F)
- Best Practices and Analytics
    - Best Practices Adoption (G) (includes most FAIR-related guidance)
    - Quality Assessed (H)
    - Analytics Alignment (I)
- Governance and Internationalization
    - Governance (good management policies and practices) (J)
    - Internationalization (K)

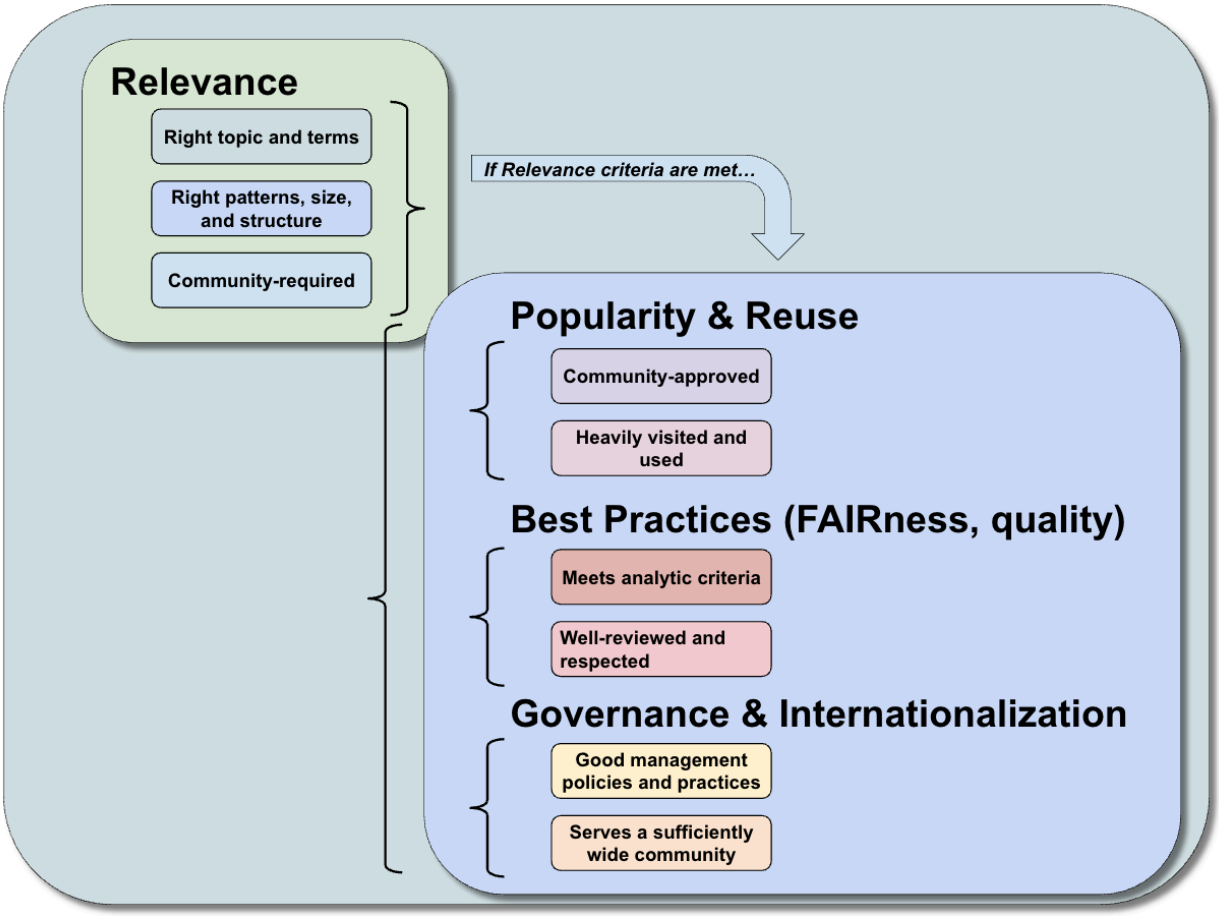


Figure aaa : One sequence of evaluations of Category Groups and Categories.

Category Descriptions and Detailed Criteria: Relevance Group

The Relevance Group contains categories that establish firm requirements for an ontology or term, such that if the requirements are not met by a particular choice, that term or ontology can not be used.

🔰 (A) Right Topics and Terms (Matching) Criteria

For ontology: the right topic and necessary metadata criteria.
Within an ontology: Matching precision, and if multiple terms, match frequency and specialization.
With respect to individual terms: level of semantic and administrative context considered in matching process; definitional checks;

Within this section, all the term-based evaluation criteria also can be used at the ontology level to determine the validity of the ontology. These uses are not specifically described, but can be algorithmically computed as a function of the totality of the term evaluations.

The Right Ontology: Topic Matching

The consideration of ontologies to address a given topic is the simplest conceptualization of a strategy for finding an ontology. The user frames the question as “What ontology can I use to address the topic of x?” The answer should be exclusionary; if a given ontology cannot address the topic x systematically to meet the user’s needs, it is rejected as not a match. Often this conceptualization is expressed using the terminology ‘semantic type’—the user wants to find an ontology matching the correct ‘semantic type’ (though this phrase typically lacks a precise definition and can cover an arbitrary range of topics). In this section we simply use the word ‘topic’, often in the form ‘major topic’ (the most high-level topics in the ontology) or ‘detailed topic’ (the topics below the top level).

Before we decompose this request into its component aspects, we note that Large Language Models have become a common means to answer questions like the one above; for details see the section Use of LLMs to Recommend Terms and Ontologies. Here we consider the atomic steps needed to address the problem reasonably rigorously, in order to identify strategies to improve the status quo (or support better LLM-based answers downstream).

An ontology may be built to address one or multiple major topics, and some or all detailed topics under the major topic(s). There are two problems to resolve if we are to evaluate the appropriateness of the ontology to satisfy a user’s ‘topical needs’: (1) how to specify the explicit terminology(ies) used to characterize the topics (of both the ontology, and the user needs); and (2) how to assign the topic—of the ontology or of the user’s interest —into the explicit terminologies used. Both of these problems have been addressed to some degree in existing ontology discovery solutions; a rough approach is outlined below.

In the first problem (specifying generally the topics of ontologies, and of user interests), many subject terminologies exist in both the library and semantic communities that could serve as ‘topic lists’, including subject terminologies in other languages. It seems reasonable to support multiple such lists, and to incorporate a map of the concepts between them, so that arriving at a ‘topic set’ for any ontology or user request is relatively interoperable. For practical implementation reasons it may be more straightforward initially to identify ‘one subject terminology to rule them all’, at least in each major domain of human knowledge, even if there is never just one terminology to rule them all. (Implementation note: These subject terminologies are often used in classification schemes (facets) to support search, as for example in FAIRsharing’s

For the second problem (assigning actual topic(s) of an ontology or user into a particular subject terminology), we first consider the assignment of an ontology’s topic(s). A topic assignment should only be made if the ontology systematically addresses that topic. Practically speaking, the full breadth of the topic should be addressed in immediately subordinate concepts, and arguably most or all of those concepts should be described in detail, ideally through another

level of concepts. The assigned subject matter and demonstrated authority level might be evaluated manually by an individual or group, or automatically through comparison of ontology content to the content of the authoritative subject terminology(ies).

The user who is looking for a ‘matching ontology’ may start with a particular topic concept, such as ‘blood cell’. If possible, the concept is directly found in, or mapped into, an authoritative subject terminology. If that is not possible (e.g, the user’s topic is too detailed), the user or algorithm must find the smallest possible broader topic in the authoritative subject terminology [e.g., ‘cell’ or ‘human cell’]. If the user has multiple classes or instances considered as “of the same type” [e.g., liver cell’, ‘tumor cell’], it is necessary for the user or algorithm to determine the common parent topic (or ontology class) for the collection, before turning to the subject terminology as just described.

Once finding the topic in the subject terminology, the search may not be complete if the topic is more narrow (specific) than any ontology topic. If that is the case, two strategies may be adopted. A search may proceed through broader topic concepts, until matching ontologies are found. Alternatively, ontology topics may be defined to an arbitrary level of detail [e.g., instead of using ‘human cell’ as the primary topic, defining the next level of cell types as topics also]. In effect, this examines increasingly lower levels of the ontology for concepts that can be treated as topics. In the most extreme case, every branch of the ontology can be considered its own topic. Note that this may be exactly what the user wants in any case: an ontology branch that addresses the topic, even one that is a small part of a much broader topic [e.g., ‘human tissue’].

A final possibility is that the topic of interest to the user is orthogonal to the topic categories in the chosen subject terminology, for example ‘3-wheeled autonomous cars’. As different facets may exist in the subject terminology for vehicle type, wheel support architecture, and vehicle control systems, users interested in ontologies or branches around highly specific topics will inevitably be better served by effective term discovery strategies.

To handle cases in which the best ontology has not been tagged with topics, we observe that searching for vocabulary lists (for example, a list of U.S. states) can often be performed easiest by searching for one or two examples (Nevada, California) and finding a shared parent concept. This is especially useful when the category has many meanings, like ‘state’.

From this reductive analysis, we may anticipate that ontology topic discovery is simply one facet of a highly optimized content search strategy, which blends term search and ontology search into a single range of search results, addressing terms, parent branches or categories, ontologies, and even ontology sets.

Simple examples of searching based on ontology topics can be found in BioPortal and AgroPortal Category facets at the left of their ontology browse pages (note the difference in level of detail), and in UMLS’ Semantic Types. Any subject-based index of content is implementing a similar concept. And while Large Language Models may provide a valuable approach to finding

ontologies based on their topics, that is addressed explicitly in the section Applicability of Large Language Models below.

The Right Ontology: Critical Metadata

The metadata describing an ontology can be an essential source for ontology qualification. Particularly for ontologies fully documented with detailed metadata, there are many attributes that can serve as discovery or exclusionary characteristics. The Metadata for Ontology Description (MOD) ontology contains the most detailed metadata specification for ontologies, although many of its attributes are too flexibly defined to allow rigorous automated assessment.

Many of the metadata attributes in MOD can address criteria described elsewhere in this document, at least to some degree. We expect the specification to evolve toward more rigorous implementation, either itself or through community-developed profiles.

If the ontology or many of its terms were derived from another ontology, metadata may also be provided (for the whole ontology and specific terms) through the PROF ontology, or via specific properties from ontologies like PROV. Community guidance for declaring the appropriate relations is being developed.

The Right Ontology: Term Match Frequency

This section and the two that follow it closely follow ontology recommendation concepts introduced in the BioPortal ontology repository, and available in all OntoPortal installations. (OntoPortal is a public domain version of BioPortal tailored for communities to offer their own ontology repository.) The BioPortal Recommender inputs a user's text selection or key topics list, and analyzes the ontologies in the repository to see how well they match the terms in the provided text or topic list.

The Term Match Frequency assesses whether the ontology matches a large number of the terms. In BioPortal Recommender this is called Coverage; other sources may call it 'term matching' or 'topic coverage'. The BioPortal algorithm[4] performs a sophisticated assessment of how much coverage is found in each ontology.

The Right Ontology: Term Match Details

The Term Match Details evaluation—in BioPortal Recommender, the 'Detail' column—determines whether the relevant terms of an ontology contain much auxiliary information (defined elsewhere as the structure measure', 'semantic richness, or granularity). (The RIght Ontology: Term Details considers whether all the terms have a desirable level of detail.)

[4] https://pmc.ncbi.nlm.nih.gov/articles/PMC5463318/

The Right Ontology: Term Match Specialization

The final BioPortal Recommender-related assessment considers whether the ontology itself is narrowly focused on the concepts being sought. If the requested terms found in the ontology make up a large part of that ontology, the Recommender infers that the ontology is built around similar topics, and the specialization is good. On the other hand, if the terms that are matched by a given ontology make up a relatively small part of the ontology, that means the ontology is covering many more topics than just the ones being matched, and so Specialization will be low.

The Right Ontology: Term Details

This section refers to the average level of detail in all the terms of an ontology. Depending on the use case, the amount of detail in an ontology's terms can disqualify the value of the ontology as a whole. Definitions are the most conspicuous concern—an ontology without many term definitions is of extremely limited value. The other annotations and relations described in The Right Term sections below can be equally important for some use cases—for example, whether an international translation exists in a particular language, or synonyms are offered to a particular reference ontology, can be a disqualifying consideration for a given community.

The Right Term: Definitions and Other Annotations

Definitions in an ontology should be supported by domain literature *or a community definition process*, and the source *or provenance* cited in the term properties. If not, the context is unclear. As a crude example, the term 'antenna' could be an anatomical part of a butterfly, or it could be a sensor on a piece of equipment. In more rigorous cases, every part of a definition that is included (or not) can affect the applicability of that term's application to another use case.

In a similar way, a term's translations, declared relations, and provenance of a term can have an impact on the value of that term for a specific use case. When this information is rigorously characterized in the ontology, automated systems can determine whether the term as a whole, or the ontology containing the terms, meets the user's criteria for acceptance.

Because term synonyms increase the ability of users to determine relevance, as well as supporting many semantic use cases more effectively, ontologies which use them should be more valued. Some use cases may even require the existence of a certain level of synonymy or synonyms from a particular semantic resource (e.g., for mappings). If the synonym is useful but not required—e.g., to assess closeness of match—we address this further in the Best Practices Adoption Criteria section.

🟢 (B) Required Ontology Structure Criteria

The Structural Criteria topic is primarily addressed in section I, Analytic Criteria. The distinction in this section is that the structure is absolutely required by the user. This is not a typical scenario, but occasionally arises in more rigorous ontology development.

For example, it may be required that the terms be incorporated within an appropriate branch hierarchy (or non-hierarchical graph), or that they be consistent with the structure of another ontology. The reader is referred to section I for more details.

🟢 (C) Community Enforced Selection Criteria

This section focuses on any firm requirements the community may establish on ontology selection and content.

Community-Required Ontology

In its simplest form, the community may simply require the use of the ontology (or certain terms) in specified ways. Such a requirement typically forces selection of the ontology or terms so required, thereby saving the user further evaluations to satisfy that use case.

When the use of the ontology or terms is performed manually—for example, by filling out values from a controlled terminology in a spreadsheet, as ImmPort submissions often do—the requirement must be validated after the fact, either manually or by a validation tool. A more expedient and increasingly common technique is to fill out values via a selection tool, as if provided with REDCap or CEDAR Workbench applications. In this case the user is presented with a list of selections that meet the criteria, and their satisfaction of the requirement is pre-ordained. (Of course the Community's representative has previously made a choice about which terms will be available as the term-selection application is configured; that evaluation is happening at the design stage for the selection tool.

Community-Required Criteria

It may also be the case that the community requires certain standards be met, or terms be provided, without prescribing use of a specific ontology. This is an indirect case of community-enforced selection criteria, in which meeting the requirement allows the ontology or term to be considered, but not meeting the requirement forces a rejection.

Because these criteria are relatively straightforward, and can directly lead to rejection of the ontology or term, we put this application of community criteria in the Relevance group. In these cases, if there is a requirement set by the evaluating community that the ontology or term doesn't meet, the ontology is by definition not relevant.

From an implementation perspective, any selection or recommendation tool may provide a choice of selection criteria with associated weights. Any such tool can and should be made more user-configurable by allowing the user to designate a criterion as Mandatory. Such a designation means that if that selection criteria is not met, the term or ontology is rejected as part of its first stage evaluation. (This can also be implemented by offering a Required Cutoff score, which will result in the ontology's rejection if the score is not met.)

Category Descriptions and Detailed Criteria: Popularity and Reuse Group

The Popularity and Reuse Group of criteria comes into play only once the Relevance Group criteria have been evaluated. For example, if the meaning of a term doesn't match the intent, or the topic of an ontology doesn't fit the need, popularity and other lesser concerns do not add any value.

If the meaning of a term or topic of the ontology is at least close to the need, but the term or ontology is not obviously the best fit, then the remaining criteria may be considered, more or less strongly according to the user's judgment of importance. For example, terms from schema.org often are too general to be particularly well matched to a specific application or community, but the extensive popularity and reuse of the schema.org standard may make it the de facto best choice.

🟢 (D) Community Approval Criteria

In the just-discussed Community Enforced Selection Criteria section, attention focuses on any firm requirements the community may establish on ontology selection. In the Governance section we discuss formal aspects of ontology and term management, focusing on defined practices that establish a norm for the ontology change process. In contrast, this section emphasizes the social roles that domain communities play in establishing and guiding ontology development, and the measurement of the ontology's visible influence on the community. (In this section we use 'domain community' or simply 'community' to refer to the people and systems that have a shared interest in the ontology.)

The OBO Foundry principles of ontology management include Principle 11: Locus of Authority. This principle declares the need for a responsible person who handles community communications and mediation, including addressing user feedback about the ontology. This informal role definition encompasses a wide range of community-related activity. Many aspects affect the construction and maintenance of the ontology to best meet the community's needs, while other aspects reflect the community's response to the ontology. These two paths of influence are presented separately.

Community impact on the ontology

The community of interest for an ontology exercises its influence in 4 main respects: (a) ensuring community knowledge is best reflected in the ontology and its terms; (b) making sure irrelevant concepts are excluded from the ontology; (c) taking into account cultural considerations of the community in the ontology's construction; and (d) providing a social network for discussion of the ontology.

Assessing the value of the ontology or its terms in relation to these community influences is subtle at best. The strongest indication of community engagement in, and impact on, an ontology is social exchanges related to the topic. Such exchanges may be in social on-line forums, whether documented (GitHub, wikis, or community websites), or in-person discussions at conferences and ontology-specific meetings or breakouts. The quantity of such exchanges, number of people participating, and maturity of the supporting infrastructure are all broad metrics of community influence on an ontology, with great significance for the maturity, breadth, and persistence of the ontology itself.

Ontology impact on the community

Several easily identified traits indicate whether the ontology (or a collection of one or more terms from an ontology) is impacting the community. Does the community endorse, suggest, or otherwise mention the ontology as an important resource? Is the ontology "built in" to community applications and services in a persistent way? Does the community have a dedicated working group or set of individuals who are tasked with influencing and/or monitoring the ontology? Are there multiple communities in the same domain working together on the ontology? Note these questions are easily answered by people, but not by automated systems, unless the answers are maintained in a consistent place and manner.

Answers to all of these questions provide strong indications as to the value of the ontology to the community, and its likelihood to be persistent and of ongoing importance. Community endorsement in particular is a highly significant indication of the value of an ontology and its terms. (Note that endorsement is a social statement by the community that an ontology is a good one; if the community actually requires an ontology or includes it in a set of approved ontologies, that is discussed in Community Enforced Selection Criteria above.)

General observation

It is assumed in evaluating this Category that the community performing the evaluation has at least a passing interest in the domain addressed by the ontology's terms. The more interested the evaluating community is in the domain of the ontology, the more important this criterion becomes.

🟢 (E) Popularity Criteria

➔ Popularity includes reuse, views, selections from a search list, and other voting techniques.

For ontology evaluation, the criterion of popularity has significant and perhaps unwarranted weight. Popularity has the value of promoting common use of terminology within the scientific community. However, it also risks elevating less important or suitable terminologies, for example promoting higher-level or more general concepts when lower-level and specific concepts are more precise. The crux of the issue is perpetuating a damaging cycle, where ontologies or terms of inferior quality gain traction and widespread use simply because of their popularity, thereby impeding the adoption of more accurate and reliable ontologies.

An illustration of this problem is seen when searching for "macrophage" in the NCBO BioPortal. The primary result is not from the Cell Ontology, despite being a domain-specific ontology tailored for cell types. The promotion of less credible assets poses a risk for users who are either unacquainted with the intricacies of ontology selection or are momentarily inattentive.

A method to counteract this concern is to measure popularity only in comparison to all other terms with similar meaning, adjusting for the popularity/adoption of the term's parent ontology. Normalization of such measurements may be challenging to perform effectively. Successful implementation also requires an assessment of 'similarity of meaning', which may be obvious when specialists consider some advanced terms, but also can be quite controversial.

Strategies for measuring popularity are diverse, but often challenging or impossible to implement for lack of good data. Even when perfectly measured, popularity only measures which terms are preferred for the participating user community; researchers with another set of requirements will not find the results helpful. For this reason popularity should be used only to a limited degree when recommending terms or ontologies.

Popularity metric: Views

➔ 'Views' refers to the number of times an ontology or term is presented to or seen by a user.

Just as Google engineers found 'page visits' alone to be an unreliable search optimization metric, most sources of ontology visit or term visit data are likely to be skewed. While for this discussion we assume that apples-to-apples comparisons are available for measuring visits to ontologies or pages, this is not broadly true across all semantic sources. Nonetheless it is measurable within some contexts, like an ontology repository with wide adoption and many ontologies and terms included.

For ontologies, two view metrics are particularly common: visits to the ontologies home page or summary page (whether in its own namespace, or within a repository), and accumulated visits to terms within the ontology. Search Engine page ranks can provide a relative metric for these

values across different presentations of all the ontologies, though ideally these would consider visits to web pages (user interface or UI) as well as application (API) calls, and could only be comparable across sites and across Search Engines if absolute page visits is the metric used. The many complexities of SEO ranking come into play when measuring visits in this way. At a repository level, the repository may be able to measure precisely the number of visits in each category (ontology and terms, UI and API).

Tool re-use of ontologies with caching may also need to be considered. Within repositories and in external tools, an ontology or term may be internally cached or indexed, so that its discovery and even use does not become visible in any measurable way. Ideally those tools should be capable of tracking and reporting the views of all the semantic content they have cached or indexed, and external applications could read and aggregate that information.

Unfortunately, it is unlikely that relatively complete view information of semantic resources will become available to the community unless there are significant monetary incentives for the repository and tool providers to provide such information.

Popularity metric: Reuse

➔ Reuse is addressed only in the Reuse Criteria section below, and not as part of popularity.

This is covered in detail in the Reuse Criteria section below. It is mentioned here because the amount of reuse can also be a factor in popularity calculations. Since this amounts to double-counting, and implementations including reuse in the popularity calculation 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

➔ 'User selection' refers to how often a term or ontology is chosen for a particular purpose.

User selection refers to the user's identification of the term or ontology as suitable for his or her purpose, for example when labelling the data or metadata of a dataset. By analogy, search engines will monitor which of the search results are clicked on, as an indication of which ones interest the user most. As with other popularity metrics, it is subject to selection bias, where the largest number of users may represent a particular community or use case, and not the inherent quality of the term or its appropriateness for a particular application.

Another fundamental difficulty in using user selection criteria for evaluating popularity is that the selection of a term or ontology happens in external tools, and is typically not made visible in a common and accessible place and via a consistent method. For a heavily used tool, with many diverse users, this statistic may prove useful in some contexts, and it would be even more useful if the statistic could be shared across all tools and contexts. As with the Views metric, this would depend on some entity incentivising the production and provision of those data.

Popularity metric: Votes and Reviews

➔ Votes and reviews are explicit opinions (ratings or text) provided by people

In the past many tools provided some way to vote on or review ontologies, in the hope that informed users could guide the selection of better ontologies. This approach has fallen out of favor, with BioPortal a particular example of an application removing reviews and voting on ontology characteristics.

In retrospect several challenges are particularly obvious:

- It takes time to perform reviews, especially systematic or credible ones, and the appropriate experts are relatively few, and are unlikely to have time to spare. So few quality reviews are created.
- Expressing one's opinion publicly and formally about a high-complexity asset is fraught, inviting unpleasant responses from those (like the asset's authors) who disagree.
- A reviewer with limited expertise will be able to add very little of value by offering an opinion of the ontology quality.
- An expert's review can only extend to that expert's use cases and understanding of the ontology; whereas users will have many other use cases and understandings that may not (often *will not*) be supported by the review.
- The credibility and quality of the ontology are less likely to be heavily considered than more fundamental qualifications (described earlier in this document).

Popularity metric: Other attention metrics

There are many other attention metrics that suggest ontology or term popularity. The number of requests (or tickets) and actual changes made to an ontology over time indicates level of engagement. The number of web searches that directly match the label or topic of the ontology can be expected to roughly track the value of the ontology. The number of projects use the ontology

The number of times one of that ontology's term identifiers is spotted 'in the wild' (in a paper, an application, or a web reference) suggests popularity, but more strongly suggests reuse, so it is not included here. Similarly the number of projects using the ontology indicates popularity, but more directly indicates reuse.

For most candidate popularity metrics not listed above, careful thought may indicate they are better considered in other categories. And while the remaining metrics can contribute to understanding popularity, they are still popularity metrics, and typically suffer from the same measurement and relevance issues described above. We see the same concerns arise in the next section, in even greater detail

(F) Reuse Criteria

Many semantic applications benefit from re-using semantic assets that already exist. In this section we review both common and subtle scenarios for evaluating this reuse. Note this is not the same as simply selecting the term for use in an external application; that is addressed in the Popularity section under User Selection. Also we point out that *indirect* reuse indications—approval or promotion by a credible authority or project(s), popularity and social metrics—are directly considered in the sections addressing those indications.

We start by acknowledging that reuse metrics, like popularity metrics, are subject to distortion effects caused by popularity of the use cases involved. The measurements must consider context to be most effective, and in many cases might be meaningful only within the same domain (or within an even more specific context). Another means of normalizing reuse measurements is to consider the specificity level of the term, and the popularity of the parent ontology, in defining a 'relative term reuse' measurement that takes into account those factors.

Perhaps the most important indicator of whether an ontology or a term is suitable for reuse is how many times the ontology or term has already been re-used. At the simplest level, some metric may exist that hints at the popularity of ontology re-use, for example 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 known 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 who need to select a subset of an ontology will create a collection of the reused statements that meet their need. 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 can help determine the value of a particular ontology or partial ontology. Where re-use specifically addresses integration with upper ontologies or parentage within other ontology models, the topic is addressed in the Best Practices subsection Integration with Upper Ontologies and External Parents.

Concerns applicable to all re-use measurements

There are several broad concerns that are significant in all types of re-use, such as being used within a context (a) that matches the desired use, (b) within a given community, (c) in a relatively current time frame, or (d) with the same meaning and precision as the desired use. These detailed concerns lead to a desire for more fine-grained or nuanced measurements of re-use—in most cases measurements that are difficult to make and rarely attempted. An analogy is searching in web search engines for *recent* information on a topic—for many searches the first references are useless due to changes in these same facets of concern (context, community, currency, meaning, and precision). Filtering on these facets would add value for many users of semantic content, but this is rarely possible. In its absence, we note that those facets correspond closely to evaluation criteria that are already described elsewhere in this document as possible primary criteria, so a significant indication of usability can be identified in those direct evaluation categories.

Measured reuse of ontology as a whole

Standards adoption is a major driver for ontology reuse, by adopting an existing ontology, implementers 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 can be considered as another type of reuse where communities build upon core concepts. If these changes are minor they are addressed through other metrics; if they are major the effect is to modify or transform an existing ontology, which we classify as whole ontology reuse. *Based on the actual use case needs, a primary base ontology may be transformed into a new ontology, including changing the hierarchies, modifying the definitions, and even changing term labels. So long as the ontology has an appropriate license, such transforms are allowed. In doing so, it's important to keep the provenance of the ontology adoption/reuse in the new ontology, in particular crediting the original ontology for its contribution.*

Measured reuse of ontology fragments, profiles, and patterns

Ontology fragments are Interrelated terms commonly used together to implicitly convey a concept or set of concepts in a domain. Examples include address (street, city, territory, country), contact information (name, email address, phone number(s)). Ontology fragments

may also be large enough to express a significant part of the source ontology; this is termed an Ontology View in the OntoPortal applications.

Ontology profiles are constrained versions of more expressive ontology language specifications or vocabularies. They are distinguished in this discussion as ‘ontology data profiles’ and ‘ontology model profiles’. Ontology data profiles, also referred to as application profiles, are collections of ontology terms derived from existing ontologies (often used to specify data exchanges, data processing, and databases). While these profiles reuse existing ontology terms, their main purpose is to provide applications and systems with verifiable and exact ontology content. Ontology model profiles are constrained versions of a more expressive ontology model, effectively defining an ontology with restricted context. Profiles can be created with explicit imports of terms and relations, or using the PROF ontology or the RDF Shape Constraint Language (SHACL) to limit the constructions offered by a particular ontology, Profiles can be published as OWL documents by standards bodies or other communities, but are also commonly published as technical documentation and behind paywalls, making their reuse metrics harder to collect by external communities.

Ontology patterns are modular fragments of ontologies designed to be used in many different domains or applications to achieve some designed goal. For example, a pattern exists for defining recurring events, and it could be used in any ontology that needs to describe recurring events. Many resources provide ontology design patterns, and these should be cited in ontologies when their patterns are reused.

The specific case addressed in all of these reuse situations is the adoption of some ontology specification from one ontological resource, and its incorporation in another semantic resource. The two principal methods for measuring this kind of reuse is discovery of the corresponding citations, either in literature or in the ontology itself (ideally described using the PROF ontology). More refined metrics and analyses may eventually support more explicit recognition of these types of reuse.

Measured 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.

Anti-patterns / false positives / improper duplication

Category Descriptions and Detailed Criteria: Best Practices and Analytics Group

🔶 (G) Best Practices Adoption Criteria

documented recommendations on ontology/term creation; includes FAIRness recommendations, use of synonyms to facilitate discovery and matching, metadata documenting the ontology. Reference 10 important practices and similar from vocabulary-building tutorial..
OBO P2) Common Format - The ontology is made available in a common formal language in an accepted concrete syntax.
*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 P6) Textual Definitions - The ontology has textual definitions for the majority of its classes and for top level terms in particular.
*OBO P7) Relations - Relations should be reused from the Relations Ontology (RO) **and/or other well-managed community ontologies 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**.*

The metric category ‘Best Practices Adoption’ looks at which generally agreed ontology best practices can be automatically detected (i.e., measured) in a given ontology. Recommended practices that must be assessed by humans are included in the Quality category.

‘General agreement’ on a practice is indicated by its inclusion in more than one recommendation, with minor differences of description or measurement ignored.

Each detailed subsection in the Best Practices category lists relevant guidance from the listed community recommendation, then discusses appropriate metrics to evaluate on ontology against the guidance.

Recommendation sources (with short name in square brackets) include:

- OBO Principles: https://obofoundry.org/principles/fp-000-summary.html
- OBO Semantic Engineering Training [OBO Training]: https://oboacademy.github.io/obook/

- AgroPortal O'FAIRe criteria [O'FAIR]: https://link.springer.com/chapter/10.1007/978-3-031-11609-4_17
- GO FAIR US FAIR criteria for ontologies and vocabularies [GFU]:
- Best Practices of Ontology Development [NIST BP]: https://www.nist.gov/system/files/documents/2021/10/14/nist-ai-rfi-cubrc_inc_002.pdf
- Handbook on Digital Engineering with Ontologies [SERC Handbook]: https://www.cto.mil/wp-content/uploads/2025/06/SERC_Handbook-on-Digital-Engineering-with-Ontologies_2.0.pdf

Computable Guidance and Recommendations

Common Format - The ontology is made available in a common formal language in an accepted concrete syntax.

FAIRness Metrics and Criteria

Common Format - The ontology is made available in a common formal language in an accepted concrete syntax.

- *The ontology is fully compliant with an established semantic web standard.*
- *The ontology provides labels and definitions for all of its concepts.*
- *The ontology is openly maintained and accessible with rich metadata in a public repository.*
- *The ontology must have a persistent resolvable identifier.*
- *The principle or best public repository for the ontology supports persistent and resolvable identifiers, versioning information, and provenance tracking for the ontology and its terms.*
- *The model defines qualified relations between its entities and integrates extensive data from multiple sources.*

Metadata Documentation

The completeness of metadata documentation can be measured against recommended or required metadata, and against the total amount of metadata provided, for whole ontologies and for specific terms. Metadata evaluations similar to that in the AgroPortal and related ontology repositories provide numerical assessments of the ontology's metadata completeness, though not of its term metadata completeness.

Providing comprehensive term metadata is a time-consuming and detailed proposition. A very few ontologies do so in detail, but this metric can be assessed by counting the average number of metadata annotations for each term.

Synonymy

Granularity and Completeness

The level of granularity in an authoritative ontology, and in any level of the ontology, should be consistent across the sub-branches or knowledge graph, to the extent the ontology represents itself as an authoritative source on that sub-topic and the lower-level content is comparably complex. Within the ontology and each of its sub-branches or topics, an authoritative ontology should be complete, by including most or all of the relevant sub-topics.

A set of metrics that provides some indication of ontology completeness might be whether that ontology is balanced at each level, so that the variance in the number of terms under each child of that level is low. Repeating that calculation across all terms that have ‘grandchildren’ would yield an indication of consistency that may prove helpful.

Integration with Upper Ontologies and External Parents

A desirable practice in developing ontologies is integrating them with appropriate parent classes, either in an upper ontology, or in another external ontology. Such integration provides the new ontology with logical and social commitments to work well with the referenced external ontology, and strengthens the power of both source and newly developed ontologies.

🔰 (H) Other Quality Assessment Criteria

➔ “Everything is quality”: Focusing on human-evaluated quality metrics that are not clearly about best practices

Many users of semantic resources want to establish ‘quality’ as a metric for evaluating ontologies and terms, and ‘insufficient quality’ is often blamed for the need to create overlapping ontological content. All of the evaluation criteria in this document can serve as quality indicators, but there is no agreement on which criteria are most important. Sometimes there is not agreement on the relationship of each criterion’s values to an assessment of quality. (For example, what is the ideal ontology size? It usually depends how you want to use the ontology.)

This section focuses on quality metrics that require human evaluation, that likely receive consistent valuations from multiple reviewers during actual assessments, that have some consensus definition, and that do not appear elsewhere in this document. In particular we exclude any of the criteria already identified as best practices in the Best Practices Adoption Criteria section.

Requirements definition and satisfaction

As with any engineered artifact, the quality of the results depends on its ability to satisfy the initial requirements for the artifact. This in turn requires that those requirements be explicitly stated. Until the requirements are clearly defined—ideally before ontology development begins— the development process will inevitably fail to achieve a coherent and satisfactory result.

For example, principle (5) from the OBO Foundry criteria explicitly states about ontology 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. To satisfy this principle, it is necessary first that the ontology have a documented set of requirements—this may take the form of requirement statements, use cases, or other written statements of the needs for the ontology."*

With this clear justification, the existence of meaningful requirements is the first measure of quality to be evaluated; the ability of the completed ontology to satisfy those requirements is the second measure. Both assessments require the judgment of semantic and domain experts.

Currency

Defined as a combination of 'recently maintained' and 'still applicable', the Currency criterion allows users to determine if the ontology is up-to-date sufficiently for the intended use, and for ongoing users, is likely to remain sufficiently up-to-date. The dominant concern is whether the ontology (hence also its terms) is in any way obsolete. While standards may be created to allow a metric calculation of this quality facet, engineering expertise is likely to remain central to this evaluation, and definition of suitable categorical values for Currency requires additional effort.

Architectural Approaches

- *Ontological realism (CUBRC p5)*
- *Multi-tiered architecture (CUBRC p6)*
- *Single-inheritance vs multi-inheritance (CUBRC p11)*
- *Content modularity/Smaller (plug-n-play) vs larger (CUBRC p8)*
- *Top-level ontology choice (BFO,*

Integration with Other Ontologies

An ontology's terms may be linked to other ontologies in 3 ways: connection to parent terms or concepts (particular to upper ontology concepts); integration and association with 'peer' concepts, especially via common properties like SKOS relation; and being connected to by lower-level terms from other ontologies. This evaluation requires some sort of measure of whether a term / branch already has been aligned with, or is designed for compatibility with, other ontologies, such that the integration can support reasoning or automated integration of data using other ontologies.

We note that it is also possible that such a connection could make the term less useful/usable, because of the complications introduced by the additional semantic constraints. The value of this metric must be determined by human evaluation of the intended application of the terms, and the suitability of the declared external integrations.

Versioning

Rigorous ontology development demands that changes be tracked and identifiers associated with each change or change set. An ideal versioning system also provides an unversioned identifier for the ontology and each term in it—this resolves to the most current released version of the ontology or term. In versioned ontologies, identifiers are needed for each version of the whole ontology, and for each version of each term in the ontology. It is also important to have access to a list of the versions, and the ability to step from one version to the next or previous version. Finally, it can be useful to look up the version of the ontology or term that applies at a given point in time.

Such a versioning approach enables detailed and automatable comparisons across any two versions of the ontology or any term in it. The access services that find and return the appropriately versioned content are provided by the repository containing the ontology. The key characteristic that the ontology must embed is a versioned identifier for its own content (kept as part of the metadata attributes of the ontology, which are declared as part of the ontology itself). With that information, any service could perform the necessary computations to provide the above versioning services.

Granularity and Completeness

Machine-computable completeness indicators may be calculated as described in the previous section, but do not verifiably reflect a valid assessment of completeness. Such an assessment requires comparison of the ontology with the current body of knowledge (as understood by domain experts, and declared by other ontologies. An ontology that contains most of the terms that appear under similar topics in other ontologies (at all levels) is likely to be granular and complete.

⭕ (I) Value-Neutral Analytic Criteria

➔ Metrics without a unidirectional value scale, but useful to compare with user goals

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 that do not serve to evaluate other Categories. 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

Simple Calculations

Percentage of terms with labels and/or definitions

Percentage of terms with synonyms or close match relations

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

Maximum number of children

Flatness (depth vs breadth)

More Complex Calculations

Need help here…

Ontology shape (?)

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

Percentage of terms with specific annotation or relation types

These metrics are similar to those specified under Simple Calculations, but with refined details that may require custom queries.

- In specific languages (also indicated by the ontology's language support metadata)
- Of specific sizes
- Using specific properties
- Relating to external terms from specific ontologies

Pattern Adoption and Other Complex Assessments

We note that the early metrics section includes a subsection on Reuse of Patterns (that is, other ontologies' reuse of patterns from the assessed ontology), whereas this section considers how the assessed ontology reuses patterns.
Need help here…

Category Descriptions and Detailed Criteria: Governance and Internationalization Group

🔰 (J) Governance Criteria

➔ Methods and control processes, trust indicators, openness of processes and participation
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.

Effective governance is also indicated by the actual operational support for the ontology during real-life dynamic events. What is the response time and response effectiveness when an issue is raised or request is made regarding that ontology? How easy and transparent is making such a request? How well does the ontology's community and governance respond when a major health emergency requires significant content or changes—can it quickly publish interim content or significant updates?

🟢 (K) Internationalization Criteria

➔ Is the resource useful across multiple languages and cultures?

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 multilingual 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 multilingual, 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.

Obvious metrics involve entities incorporating multi-language annotations. Considerations include what number and percentage of existing annotation types (label, description, note, etc.) have multilingual annotations, the number of languages support, and possibly weighting higher-level (broader) terms more when calculating the value of multi-language annotations.

Credibility of Translations

Multiple factors affect the credibility of translated content. First is the question of how that non-primary-language content was generated—it could be manually generated, by an individual or a community (better). It could also be automated by any number of tools, with different levels of credibility. Tools using large language models have to date proven themselves needy of expert curation, especially for highly specialized terms like those in the medical domain.

The level of metadata provided with the translation is also a credibility indicator. Metadata indicating when, how, and by whom a translation was created should be included with each term.

Whether any manual or post-translation curation or validation of the translated information is also a strong indication of credibility.

Finally, an indication of the recency of the translation is valuable, though this can be tricky to measure precisely—lack of an update to a translation does not strongly correlate with less validity. The processes used to produce the international content are important considerations affecting the relevance of the ‘last update’ time.

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 somewhat unavoidable in the domain name, it is more avoidable in the rest of the namespace, and entirely 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?

Existing Evaluation Systems, Technologies, and Models

🔶 Existing Tools

A list and description of tools/technologies that claim to evaluate terms or ontologies. List the tool or technology, what category it evaluates (term, ontology, or both), a link to the tool or technology if available, and a very brief description.

To describe a particular tool or technology with 1 to 2 paragraphs, create a subheading below the table. Focus on the features of the resource that support evaluation of terms or ontologies. (Tools with implicit selection of terms, like mapping tools or annotation tools, may also be included. Describe their primary purpose and how they implement term evaluation.)

Tool or Technology	***Target (Term, Ontology, Both)***	***Link or Reference***	***Brief description***
BioPortal	*Both*	*bioportal.bioontology.org*	*Large collection of community ontologies for biomedicine and more (n.b. Recommender, Annotator)*
AgroPortal O'FAIRe	*Ontology*	*agroportal.lirmm.edu*	
OBO Foundry Ontology Browser	*Ontology*	*https://obofoundry.org/*	*Curated list of coordinated biomedical ontologies with status dashboard*
BODC Semantic Analyzer	*Both*	*https://semantics.bodc.ac.uk/*	*To be provided (from Alexandra's presentation)*
EMBL-EBI Ontology Lookup Service	*Both*	*https://www.ebi.ac.uk/ols4/*	*Allows searching for terms or browsing a filtered ontology list from a curated list of biomedical ontologies*
OntoFox			
OOPS			
FOOPS			
CEDAR Workbench	*Both*	*cedar.metadatacenter.org*	*Metadata template creator that lets users choose what ontologies, branches, or terms to use in templates and instances*

BioPortal Ontology Repository Recommender&Annotator

AgroPortal Ontology Repository FAIR Evaluator

OBO Foundry Ontology Browser

BODC Semantic Analyzer

CEDAR Workbench

🔶 Implemented Technologies

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

Term Evaluations

- value of enclosing ontology [BioPortal, OntoPortal]
- syntactic match [BioPortal Annotator/Annotator+, OntoPortal]
- identifier match [BioPortal, OntoPortal]
- desired domain (can come from enclosing ontology)
- Existing (documented) matches or mappings, through annotations, mappings, etc.)

Ontology Evaluations

- Ontology Recommender (version 2) [BioPortal, OntoPortal]
    - match frequency
    - specialization of ontology
    - popularity (as measured locally)
- Evaluation criteria dashboards [OBO Foundry]
- Level of reuse [internal in BioPortal, OntoPortal; implied in GitHub-based ontologies]
- Ontology reviews [AgroPortal, formerly BioPortal]
- Metadata adoption and FAIR evaluators
    - O'FAIRe [AgroPortal]

🔶 Published Approaches (Not Yet Implemented)

Term Evaluations

- Natural Language and Text Annotation techniques
- Machine Learning techniques

- Large Language Model techniques

Ontology Evaluations

-

Applicability of Large Language Models (LLMs)

This is becoming relevant enough to write some words about it. The challenge is to write only the necessary words for the user to evaluate whether LLMs will be helpful in any of these ways, particularly whether/how they can lead to better choices for terms and ontologies. This needs to be smaller than a review article, and more focused on explaining the status in each category than closely analyzing the details of these processes.

⭕ Ability of LLMs to Process and Explain Existing Terms and Ontologies

To my knowledge currently LLMs' capability to 'understand' (use) an ontology in its native format is low, but if a translator converts the ontology to a human-readable description of the contents, the LLM can use the ontology very well. This section should confirm or deny the level and methods of LLM ontology processing and explanation.

⭕ Use of LLMs to Recommend Terms and Ontologies

At least one anecdotal report suggests LLMs can be very convincing in recommending terms and ontologies for particular purposes. Probably accuracy is variable, because LLMs; but as a strategy available to non-expert users this could be a powerful and/or tempting shortcut. It would be helpful to be authoritative here, with many citations.

⭕ Pre-conditioning LLMs With Existing Schema Before Recommending

Often an LLM is able to make much better outputs if it is 'trained' with information about the starting context (e.g., the context into which ontologies and their terms are inserted), and maybe about the options under consideration (if not 'all ontologies everywhere').

⭕ Generating Ontology Content Directly with LLMs

I claim this topic doesn't belong in this white paper. The only connection is if we want to talk about the fact that the generation process may include choosing (or reproducing) some existing terms or ontologies as part of the generated content. Which raises a whole 'nother slew of issues.

Recommended 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 and Their Impact

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, briefly discussing the impact of each use case on the selection criteria.

Following the use case list, we include a table highlighting the estimated weight of each criterion in addressing a particular use case. This Impact Heat Map provides a template for weighting the different criteria when considering a specific use case.

🔶 Searching for a term to…

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

Reuse in an AI application

Add parts to your own ontology

🔶 Searching for a set of terms

Name attributes to describe an object

Characterize current state of research

Example of this: work to define current research in a particular field like acoustic measurements of marine biological species

🔶 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

Finding ontologies that address a particular topic ('semantic type').

🔶 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)

🔶 Sources for additional use cases

The following organizations or projects may provide use cases that are particularly well defined or specialized relative to the above list.

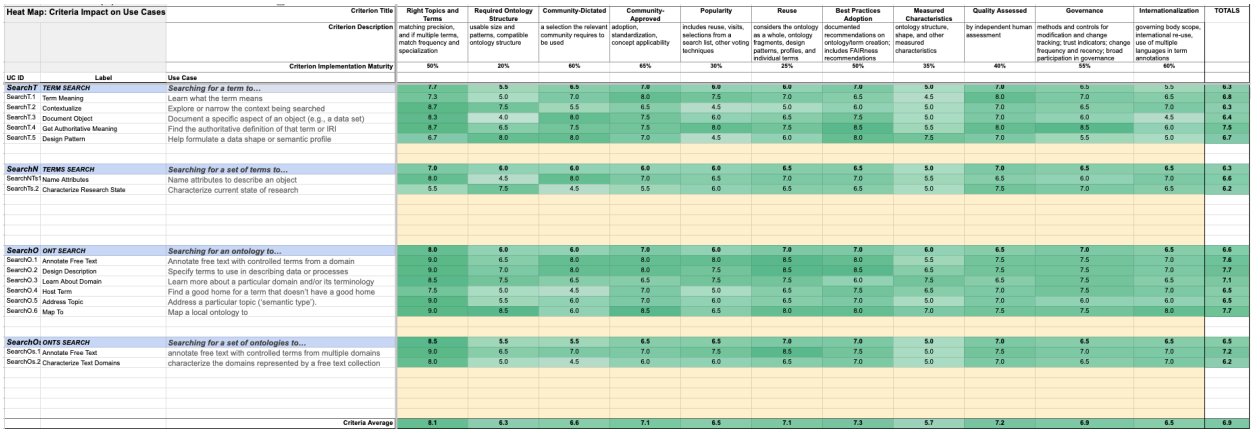
- I-ADOPT would like to be able to target application ontologies that satisfy their abstract ontology needs.
- The Environmental Health Language Consortium at NIH/NIEHS wants to create a list of terms that can be used in common across the division or all NIH.

🔶 Use Case Impact Heat Map

This visualization (copied from Heat Map: Criteria Impact on Use Cases) illustrates the relevance of each evaluation criterion for different choosing terms and ontologies' use cases. A user could choose to weight the importance of the evaluation criteria above to reflect the values shown in the heat map.

In the table, the use cases above are shown down the left-hand side of the table. The top row describes the criterion being considered for that use case.

The number and color in each cell illustrates the importance of the relevant criterion for a given use case. The importance is given as a number from 0 to 10, with 0 shown as white, and larger numbers shown in increasingly dark shades of green.



Heat Map: Criteria Impact on Use Cases		Criterion Title	Right Topics and Terms	Required Ontology Structure	Community-Dictated	Community-Approved	Popularity	Reuse	Best Practices Adoption	Measured Characteristics	Quality Assessed	Governance	Internationalization	TOTALS
		Criterion Description	matching precision, and if multiple terms, match frequency and specialization	usable size and patterns, compatible ontology structure	a selection the relevant community requires to be used	adoption, standardization, concept applicability	includes reuse, visits, selections from a search list, other voting techniques	considers the ontology as a whole, ontology fragments, design patterns, profiles, and individual terms	documented recommendations on ontology/term creation; includes FAIRness recommendations	ontology structure, shape, and other measured characteristics	by independent human assessment	methods and controls for modification and change tracking; trust indicators; change frequency and recency; broad participation in governance	governing body scope, international re-use, use of multiple languages in term annotations	
		Criterion Implementation Maturity	50%	20%	60%	65%	30%	25%	50%	35%	40%	55%	60%	
UC ID	Label	Use Case												
SearchT	***TERM SEARCH***	***Searching for a term to…***	7.7	5.5	6.5	7.0	6.0	6.0	7.0	5.0	7.0	6.5	6.5	6.3
SearchT.1	Term Meaning	Learn what the term means	7.3	5.0	7.0	8.0	7.5	7.0	6.5	4.5	8.0	7.0	6.5	6.8
SearchT.2	Contextualize	Explore or narrow the context being searched	8.7	7.5	5.5	6.5	4.5	5.0	6.0	5.0	7.0	6.5	7.0	6.3
SearchT.3	Document Object	Document a specific aspect of an object (e.g., a data set)	8.3	4.0	8.0	7.5	6.0	6.5	7.5	5.0	7.0	6.0	4.5	6.4
SearchT.4	Get Authoritative Meaning	Find the authoritative definition of that term or IRI	8.7	6.5	7.5	7.5	8.0	7.5	8.5	5.5	8.0	8.5	6.0	7.5
SearchT.5	Design Pattern	Help formulate a data shape or semantic profile	6.7	8.0	8.0	7.0	4.5	6.0	8.0	7.5	7.0	5.5	5.0	6.7
SearchN	***TERMS SEARCH***	***Searching for a set of terms to…***	7.0	6.0	6.0	6.0	6.0	6.5	6.5	5.0	7.0	6.5	6.5	6.3
SearchNTs1	Name Attributes	Name attributes to describe an object	8.0	4.5	8.0	7.0	6.5	7.0	7.0	5.5	6.5	6.0	7.0	6.6
SearchTs.2	Characterize Research State	Characterize current state of research	5.5	7.5	4.5	5.5	6.0	6.5	6.5	5.0	7.5	7.0	6.5	6.2
SearchO	***ONT SEARCH***	***Searching for an ontology to…***	8.0	6.0	6.0	7.0	6.0	7.0	7.0	6.0	6.5	7.0	6.5	6.6
SearchO.1	Annotate Free Text	Annotate free text with controlled terms from a domain	9.0	6.5	8.0	8.0	8.0	8.5	8.0	5.5	7.5	7.5	7.0	7.6
SearchO.2	Design Description	Specify terms to use in describing data or processes	9.0	7.0	8.0	8.0	7.5	8.5	8.5	6.5	7.5	7.5	7.0	7.7
SearchO.3	Learn About Domain	Learn more about a particular domain and/or its terminology	8.5	7.5	6.5	6.5	7.5	7.5	6.0	7.5	6.5	7.5	6.5	7.1
SearchO.4	Host Term	Find a good home for a term that doesn't have a good home	7.5	5.0	4.5	7.0	5.0	6.5	7.5	6.5	7.0	7.5	7.0	6.5
SearchO.5	Address Topic	Address a particular topic ('semantic type').	9.0	5.5	6.0	7.0	6.0	6.5	7.0	5.0	7.0	6.0	6.0	6.5
SearchO.6	Map To	Map a local ontology to	9.0	8.5	6.0	8.5	6.5	8.0	8.0	7.0	7.5	7.5	8.0	7.7
SearchOs	***ONTS SEARCH***	***Searching for a set of ontologies to…***	8.5	5.5	5.5	6.5	6.5	7.0	7.0	5.0	7.0	6.5	6.5	6.5
SearchOs.1	Annotate Free Text	annotate free text with controlled terms from multiple domains	9.0	6.5	7.0	7.0	7.5	8.5	7.5	5.0	7.5	7.0	7.0	7.2
SearchOs.2	Characterize Text Domains	characterize the domains represented by a free text collection	8.0	5.0	4.5	6.0	6.0	6.5	7.0	5.0	7.0	6.5	7.0	6.2
		Criteria Average	8.1	6.3	6.6	7.1	6.5	7.1	7.3	5.7	7.2	6.9	6.5	6.9

Evaluation Guidance

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

Although the semantic selection landscape has many complexities, users need applicable guidance to meet their own semantic needs. This section offers our conclusions about the best way(s) to proceed with a semantic selection process.

As a start, users must be able to find resources that are possibly applicable to their semantic needs. We did not set out to address that need, but our Discovery Paths section here provides references to existing guidance on the topic.

🔰 Discovery Paths

The references below offer guidance to help you discover candidate ontologies and terms. The three Selection sections (that follow these references) consolidate the knowledge in this document to some specific all-purpose recommendations for choosing the best ontologies and terms from your discovered options.

- *(pending)*
- *BARTOC*
- *BioPortal/OntoPortal*
- *Linked Open Vocabularies (LOV) [lots of bad links and links to bad ontologies though]*

🔰 Selection Now: By Yourself With Current Tools

Note OBO Academy references

The selection paths available to you today are limited by the selection tools and metrics that are implemented. Our guidance focuses on characteristics that can be manually evaluated—even by an inexperienced individual—and are likely to yield a positive result. Where existing tools can be helpful, we explain their advantages and weaknesses.

Of course, if your discovery process has only yielded one candidate ontology or term, you can spare yourself further ‘selection’ evaluations. Your only concern is whether your discovered resource is adequate. Here again the criteria below may be useful, but if you realize your asset is not sufficient, we will not tell you in this document how to build your own ontology or vocabulary to satisfy your needs. There are many documents on the web describing this process, for all levels of user experience.

Most Important Criteria

As the rubric in the Typical Evaluation Criteria section suggests, primary criteria for almost every use case are in the Relevance category: the right ontology and term types, and any rigid requirements you may have to satisfy your use case. So at a basic level, answer the question “does this semantic asset meet the must-have requirements?” of meaning, function, and community rules.

Beyond those basics, the highest value criteria for ontologies and terms are community recommendations, adoption of best practices, and assessed quality. Of these, assessed quality is the most subjectively appraised, and often amounts to “ask an expert (or someone who knows more than you do)”. Since assessed quality is often strongly correlated with community recommendation and adoption of best practices, and it is not readily discoverable for most ontologies, we won’t consider it further in this section.

Of course other specific needs may factor into your decision, and you can add and evaluate criteria like internationalization as needed to maximize your goals. For the remaining two criteria, Community Recommended and Best Practices Adoption, here are a few straightforward strategies to perform your evaluation.

Community Recommended: Which assets are already adopted (or even developed) by your domain or project? You can ask community members to learn what other people are using—this is usually a strong indicator of a suitable selection. Some communities, like Elixir, have

produced cookbooks[5] that describe how specific terms or ontologies can be found and selected. https://faircookbook.elixir-europe.org/content/recipes/interoperability/selecting-ontologies.html

Best Practices Adoption: While this may seem a bit fuzzy, the Available Tools section offers some specific references and services for evaluating good ontology practices. Even if your ontology isn't evaluated by those tools, you can often use your judgment to compare how it ranks in FAIRness, metadata quality, and meeting basic principles of ontology and term creation.

Finally, if these results do not clarify the best choice for your needs (and even if they do), it is important to consider how the ontology or term will function in your application. Perform a 'dry run' by mentally evaluating how the semantic information will be used in your use case. If it seems like a good fit in that context, you can probably move forward with the expectation that it will likely serve you well.

Available Tools

BioPortal: Search (terms), Recommender (ontologies and terms), Annotator (terms and ontologies)
OntoPortal deployments like AgroPortal: O'FAIR Evaluator (best practices), metadata completeness (best practices)
OBO Library: ontology badges indicate best practices adoption
FAIRsharing: community adoption
LLMs: is a bit of a hail mary in terms of credibility, but can provide useful hints and often astonishingly good advice
OLS (ontology lookup service)
OntoFox
NVS tooling (Semantic Analyzer)

Manual Evaluation

(depending on your use case and situation)
Some of the additional categories
Easiest to get expert advice to weigh different options
Walk through each category briefly (?)

🔶 Selection Future: In A Better System

More metrics will exist for many of these criteria
Popularity and Reuse will be at least somewhat rigorously measured
More tools will perform more wide-ranging evaluations
Example: Aryan's tool

[5] Selecting terminologies and ontologies, https://faircookbook.elixir-europe.org/content/recipes/interoperability/selecting-ontologies.html

LLMs will be able to provide (somewhat) meaningful feedback around your particular needs

🔶 Selection Far Future: In An Ideal System

Ontology creation process is advanced and adopts most best practices routinely
All ontologies are indexed in systems designed to support comparisons (in addition to their favority repositories), allowing ready comparison
LLMs or their next incarnation will be able to detect your needs and use their knowledge of the semantic resources (or of the tools that characterize the semantic resources, or both) to make recommendations to your specific use cases

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 and Bibliography

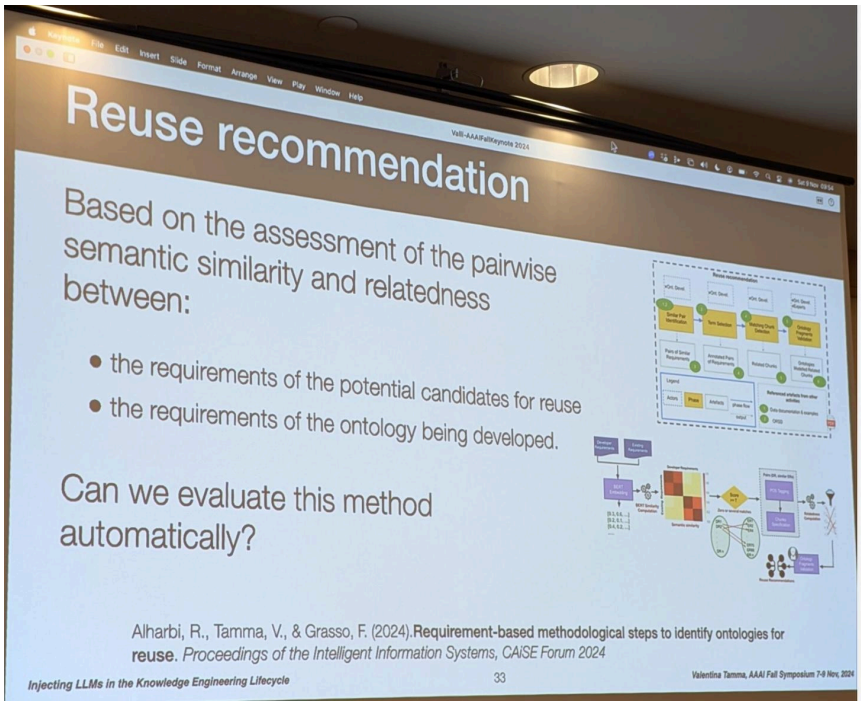
Many references related to ontology and term creation and selection have been collected at *Vocabulary References focusing on SKOS*. *That document attempts to bring to the top the most useful advice and tools for the early-stage user of semantic resources, and also references more sophisticated content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

Please cite existing publications and resources that can be used to advance the evaluation of terms and ontologies. Highlight those resources that are particularly accessible.

Resources

Bibliography



See endnote [1]

[1] Ahlarbi, R., TRamma, V., & Grasso, F. (2024) **Requirement-based methodological steps to identify ontologies for reuse**. Proceedings of the Intelligent Information systems, CAISE Forum 2024.

- https://faircookbook.elixir-europe.org/content/recipes/interoperability/selecting-ontologies.html (review to see if we have it covered)