

Adapting the Harmonized Data Quality Framework for Ontology Quality Assessment

Tiffany J. Callahan^{1,2}, William A. Baumgartner Jr², Nicolas A. Matentzoglou³, Nicole A. Vasilevsky¹, Lawrence E. Hunter², Michael G. Kahn²

¹University of Columbia, ²University of Colorado Anschutz Medical Campus, ³Semanticly Ltd

ABSTRACT

Motivation: Ontologies play an important role in the representation, standardization, and integration of biomedical data, but are known to have data quality (DQ) issues. We aimed to understand if the Harmonized Data Quality Framework (HDQF), developed to standardize electronic health record DQ assessment strategies, could be used to improve ontology quality assessment. A novel set of 14 ontology checks was developed. These DQ checks were aligned to the HDQF and examined by HDQF developers. The ontology checks were evaluated using 11 Open Biomedical Ontology Foundry ontologies. 85.7% of the ontology checks were successfully aligned to at least 1 HDQF category. Accommodating the unmapped DQ checks (n=2), required modifying an original HDQF category and adding a new Data Dependency category. While all of the ontology checks were mapped to an HDQF category, not all HDQF categories were represented by an ontology check presenting opportunities to strategically develop new ontology checks. The HDQF is a valuable resource and this work demonstrates its ability to categorize ontology quality assessment strategies.

1 INTRODUCTION

Ontologies play an important role in the representation, standardization, and integration of knowledge [1]. Much like electronic health record (EHR) data, ontologies are subject to quality issues which can be difficult to assess. Typically, quality assessment is performed during the development process and includes the identification of logical or semantic (i.e., inconsistency, incompleteness, and redundancy), syntactic, and/or hierarchical errors at the schema- (i.e., intrinsic structural evaluation) and data-level (i.e., instances) [2]. A wide-range of strategies have been developed to detect these types of errors through the application of metrics like consistency, interpretability, accuracy, and relevance [3]. Unfortunately, existing tools can be difficult to use, are inconsistently maintained, and may not be publicly available [4,5]. The majority of ontology quality assessment resources have been limited to a single ontology, but serious issues can arise when combining multiple ontologies or within the context of a specific use case [6]. Thus, additional strategies to assess the quality and fitness of multiple ontologies are needed.

EHR “fitness-of-use” assessment has been exhaustively researched [7]. Similar to ontology evaluation, many quality assessment strategies and metrics have been developed [8], but until recently, there was little consensus on which dimensions mattered most and how to best assess them. The Harmonized Data Quality Framework (HDQF) [7] was developed to solve this problem. The HDQF characterizes data quality (DQ) assessment strategies and metrics, allowing existing strategies, metrics, and checks implemented and documented in different ways, from different organizations, to be compared. The framework has two evaluation contexts; confirmation of expectations based on comparisons to local knowledge (*Verification*) and external benchmarks (*Validation*). Within these contexts, there are 3 dimensions and 8 categories:

1. **Conformance.** Values adhere to formatting (*Value*; e.g., sex only has values “Male”, “Female”, or “Unknown”) and structural (*Relational*; e.g., patient identifiers link data from different tables) constraints and are accurate when computationally derived (*Calculation*; e.g., database- and hand-calculated BMI values are identical).
2. **Completeness.** Values are present at a single (*Atemporal*; e.g., gender is not NULL) or multiple (*Temporal*; e.g., discharge dates are not missing for consecutive days) point(s) in time.
3. **Plausibility.** Values are believable when assessed through the agreement of independent (*Atemporal*; e.g., height has a positive value) or temporal measurements (*Temporal*; e.g., date of birth occurs before date of death) of the same fact and are not duplicated (*Uniqueness*; e.g., concepts are not referenced by multiple identifiers).

The HDQF is widely used by the clinical community and may provide an avenue for expanding existing ontology quality assessment strategies to multiple ontologies and/or the fitness of ontologies to address specific use cases. The goal

of this work was two-fold: (i) develop and evaluate ontology checks to assess the fitness-of-use of individual and merged biomedical ontologies and (ii) characterize these checks using the HDQF. An expanded version of this work is available: <https://doi.org/10.5281/zenodo.5716401> (Section 2.2.1).

2 METHODS

2.1 Ontology Check Development

The ontology checks were developed using a Python 3.6.2 Jupyter Notebook. The ontology checks were reviewed by a member of the Open Biomedical Ontology (OBO) Foundry Working Group and by a Semantic Web developer with >10 years of experience (11/2020-03/2021). For evaluation, the ontology checks were applied to 11 of the most frequently used OBO Foundry ontologies: Chemical Entities of Biological Interest (ChEBI), Cell Line Ontology (CLO), Gene Ontology (GO), Human Phenotype Ontology (HP), Mondo Disease Ontology (Mondo), Protein Ontology (PR), Pathway Ontology (PW), Relation Ontology (RO), Sequence Ontology (SO), Uberon Multi-species Anatomy Ontology (UBERON), and Vaccine Ontology (VO). Results were output to an ontology quality report that includes statistics pre/post ontology checks.

2.2 Ontology Check Characterization

The original published version of the HDQF [7] was used for the ontology check alignment. The ontology checks were mapped using two contexts (i) a single ontology and (ii) a set of merged ontologies. The original HDQF developers independently mapped each of the ontology checks to the HDQF. Any check not able to be clearly mapped was discussed until a final mapping consensus was reached. All code, Notebooks, and data are publicly available (<https://zenodo.org/record/6468948>).

3 RESULTS

The developed set of ontology checks and definitions are shown in Table 1. Experts provided feedback regarding the usefulness of each DQ check, definition, algorithm solutions to address each error, and the format and content of the ontology quality report. Two iterations of revisions were required to finalize the check definitions, 3 iterations of revisions were required to finalize the current version of the algorithmic solutions, and 4 iterations of revisions were required to create the ontology quality report (see Zenodo link for example).

The ontology checks were applied to 11 OBO Foundry ontologies. The CLO had the most errors (i.e., 1 *Value Error*, 16 *Punning Errors*, 13 *Obsolete Entities*, and 2 *Deprecated Entities*) and VO had the fewest (2 *Identifier Errors*). The amount of deprecated entities ranged widely across the ontologies with ChEBI (n=18,506), SO (n=341), and GO (n=6,430) containing the greatest proportion with respect to the total number of classes. *Punning Errors* were only identified in the CLO (n=16) and when merging the 11 ontologies (n=8). All of the individual ontologies were deemed to be consistent via the ELK reasoner, which was run after the DQ checks were performed. *Semantic Heterogeneity* errors were only detected in the set of merged ontologies (7 native ontology classes and 23,624 imported classes required identifier normalization).

The HDQF categories aligned to each DQ check within the individual and merged ontology contexts are shown in Table 1. All of the ontology checks except *Deprecated* and *Obsolete Entities* were mapped to an HDQF category which led to the development of a new *Data Dependency* category to account for data versioning. When aligning the ontology checks within the context of an individual ontology to the HDQF, 42.8% mapped to *Value Conformance* (i.e., *Value Errors*, *Identifier Alignment*, and *Identifier Errors*), 28.6% to *Plausibility Uniqueness* (i.e., *Semantic Heterogeneity* and *Punning Errors*), and 28.6% to *Data Dependency* (i.e., *Deprecated* and *Obsolete Entities*). Within the merged ontology context, all of the ontology checks mapped to *Relational Conformance*.

Table 1. Ontologies Data Quality Checks.

Check	HDQF Alignment	Check Description
Deprecated Entities	^a Data Dependency ^b Relational Conformance	These checks identify entities within an ontology that have been staged for removal or marked as no longer in use. The presence of staged entities in an ontology is expected given the evolution of these resources across the data cycle. These specific checks were developed to support the use of ontologies within scenarios where the inclusion of staged entities has the potential to introduce opportunities for downstream errors. For example, introducing staged entity dependencies in newly constructed classes and axioms requires the maintenance of metadata regarding the status of these entities for all downstream applications.
Obsolete Entities	^a Data Dependency ^b Relational Conformance	All deprecated and obsolete entities as well as the triples that utilize them are removed. For each ontology, removed entities are organized by status (i.e., obsolete versus deprecated) and entity type (i.e., owl:Class vs. owl:ObjectProperty) in the ontology quality report.
Value Errors	^a Value Conformance ^b Relational Conformance	This check utilizes logic from the Owlready2 Python library to identify errors related to invalid typing of literals and owl:Class, owl:ObjectProperty, owl:AnnotationProperty, and owl:NamedIndividual entities. CLO contained the error "an Invalid literal for int() with base 10". The erroneously defined literal was re-typed. Issue: https://github.com/CLO-ontology/CLO/issues/48
Semantic Heterogeneity	^a Plausibility Uniqueness ^b Relational Conformance	This check looks for entities that represent the same concept, but which are referenced by different identifiers and built using different definitions. This check logically aligns entities identified as the same by making them <code>rdfs:SubClassOf</code> a single entity from the ontology considered to be the primary for that domain. For example, multiple occurrences of the following concepts were independently defined: "gene" (SO and the VO), "protein" (ChEBI, PR, and SO), "Disorder" (VO and MONDO), "antigen" (VO and the ChEBI), "gelatin" (VO and the ChEBI), and "hormone" (VO and ChEBI). To address this, the check takes as input a set of instructions for how to normalize duplicated entities. The SO was used as the primary entity for all "gene" and "protein" occurrences, MONDO was used as the primary entity for all "disorder" occurrences, and ChEBI was used for all "antigen", "gelatin", and "hormone" occurrences.
Identifier Alignment	^a Value Conformance ^b Relational Conformance	This check identifies errors or inconsistencies in the identifiers of imported entities that are not part of an ontology namespace. This check requires a reference dictionary to verify the syntax and status (e.g., updated or withdrawn) of identifiers. A withdrawn entity was discovered when converting PR-imported gene identifiers from HGNC to Entrez Gene. Issue: https://github.com/PROconsortium/PROteinOntology/issues/176
Identifier Errors	^a Value Conformance ^b Relational Conformance	This check identifies syntax errors in the formatting of ontology identifiers. Currently, this check identifies namespace errors in compact URIs or CURIEs (Compact Uniform Resource Identifier). For example, the incorrect use of <code>HPO_</code> instead of <code>HP_</code> . It can be challenging to identify these types of errors within a single ontology without utilizing an external gold standard or reference resource to verify identifiers against (e.g., identifying a primary ontology for each domain and requiring the identifiers of all other ontologies that reference it to match). Erroneous identifiers are repaired via the use of a reference resource. In the VO, an incorrectly formatted occurrence of <code>PRO_XXXXXX</code> was replaced with <code>PR_XXXXXX</code> . Issue: https://github.com/vaccineontology/VO/issues/4
Punning Errors	^a Plausibility Uniqueness ^b Relational Conformance	This check identifies entities with more than one type declaration. This check follows recommendations of Vrandeic (2010) [9], such that all entities are checked for redeclarations and when identified, the following corrections are made: (i) Entities typed as an <code>owl:Class</code> and an <code>owl:NamedIndividual</code> are typed as an <code>owl:Class</code> (ii) Entities typed as an <code>owl:ObjectProperty</code> and an <code>owl:AnnotationProperty</code> are typed as an <code>owl:ObjectProperty</code> This check identified 7 illegally punning errors in the CLO. Issue: https://github.com/CLO-ontology/CLO/issues/43

^aApplied to an Individual ontology; ^bApplied to merged ontologies. See code for each check here: https://github.com/callahan/tj/PheKnowlator/blob/master/notebooks/Ontology_Cleaning.ipynb

4 CONCLUSIONS

A total of 14 ontology checks were developed and successfully mapped to the HDQF when it was extended to include a new *Data Dependency* category. The *Data Dependency* category was created to address dependency and versioning errors. While all of the ontology checks were mapped to an HDQF category, not all HDQF categories were represented by an ontology check (i.e., *Relational Calculation*, *Atemporal/Temporal Completeness*, and *Atemporal/Temporal Plausibility*). These categories can be prioritized and/or help guide new ontology check development. Comparing these results to the characterization of EHR data DQ checks performed in our prior work [8], resulted in some important differences; most EHR DQ checks mapped to *Temporal Plausibility* and *Atemporal Completeness* whereas most ontology checks mapped to *Relational Conformance*. This makes sense, the plausibility and completeness of clinical data are important indicators of its secondary use. With respect to ontologies, whether or not an ontology can be combined with other ontologies is very important for complex tasks like building knowledge graphs. The fact that the HDQF could characterize clinical and ontology checks provides initial support for its ability to characterize the quality assessment of translational resources.

Standardized processes or principles help create high-quality ontologies by enabling a more reproducible and consistent development environment [2]. While the use of principles to guide the development of ontologies is important, many independent sets of principles have been developed, which can be overwhelming and confusing for practitioners, and without alignment, can decrease intra-ontology interoperability [1]. Among open-source ontologies, the specific principles used during development are not always explicitly described and it may be unclear as to whether ontologies are kept up-to-date with changes in principles over time. The OBO Foundry is one of the largest collaboratives dedicated to establishing robust tools in support of ontology development. A dedicated community, the OBO Foundry is constantly improving existing tools and developing novel solutions to improve ontology quality. Their recent article [10], introduced a dashboard to log the status of each ontology with respect to their 13 principles [11]. None of the principles addressed our ontology checks.

This work has limitations. First, the ontology checks were tested on a small sample of ontologies. Second, we have not yet mapped existing ontology checks, like those underlying the OBO dashboard, to the HDQF. Third, our current ontology quality pipeline does not include existing tools like ROBOT [12]. Finally, ontology developers did not actively participate in the HDQF alignment process. Obtaining additional feedback on the checks and the HDQF alignment could generate ideas for additional ontology checks, including the HDQF categories without mapped checks.

REFERENCES

- Blake JA, Bult CJ. Beyond the data deluge: data integration and bio-ontologies. *J Biomed Inform.* 2006;3:314–20
- Köhler J, Munn K, Rüegg A, et al. Quality control for terms and definitions in ontologies and taxonomies. *BMC Bioinformatics.* 2006;7:212
- Brank J, Grobelnik M, Mladenic D. A survey of ontology evaluation techniques. Proceedings of the conference on data mining and data warehouses. 2005;166–70
- Amith M, He Z, Bian J, et al. Assessing the practice of biomedical ontology evaluation: Gaps and opportunities. *J Biomed Inform.* 2018;80:1–13.
- Parsia B, Sirin E, Kalyanpur A. Debugging OWL ontologies. Proceedings of the 14th international WWW conference. 2005;633–40
- Slater LT, Gkoutos GV, Hoehndorf R. Towards semantic interoperability: finding and repairing hidden contradictions in biomedical ontologies. *BMC Med Inform Decis Mak.* 2020;20(Suppl 10):311
- Kahn MG, Callahan TJ, Barnard J, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *eGEMS.* 2016;4:1244
- Callahan TJ, Bauck AE, Bertoch D, et al. A Comparison of Data Quality Assessment Checks in Six Data Sharing Networks. *eGEMS.* 2017;5:8.
- Vrandeic Z. Ontology Evaluation. Karlsruhe Institut für Technologie; 2010
- Jackson RC, Matentzoglou N, Overton JA, et al. OBO Foundry in 2021: Operationalizing Open Data Principles to Evaluate Ontologies. *bioRxiv.* 2021
- OBO Working Group. <http://obofoundry.org/principles/fp-000-summary.html>
- Jackson RC, Balhoff JP, Douglass E, et al. ROBOT: A Tool for Automating Ontology Workflows. *BMC Bioinformatics.* 2019;20:407