

# Adapting the Harmonized Data Quality Framework for Ontology Quality Assessment



Tiffany J. Callahan<sup>1,2</sup>, William A. Baumgartner Jr<sup>1</sup>, Nicolas A. Matentzoglou<sup>3</sup>,  
Nicole A. Vasilevsky<sup>1</sup>, Lawrence E. Hunter<sup>1</sup>, Michael G. Kahn<sup>1</sup>

<sup>1</sup>Computational Bioscience Program, University of Colorado AMC;  
<sup>2</sup>Department of Biomedical Informatics, Columbia University; <sup>3</sup>Semanticly Ltd



## BACKGROUND

- Ontologies play an important role in the standardization and integration of knowledge and are subject to data quality issues.<sup>1</sup>
- Existing ontology quality assessment tools can be difficult to use, are inconsistently maintained, and may not be publicly available.<sup>2-5</sup> Most tools evaluate a single ontology; serious issues arise when using multiple ontologies or within the context of a specific use case.<sup>6</sup>
- Many clinical data quality assessment strategies exist,<sup>8</sup> but until recently, there was little consensus on what mattered most and how to best assess it. The Harmonized Data Quality Framework (HDQF)<sup>7</sup> was developed to solve this problem.

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.

## METHODS

### Ontology Check Development

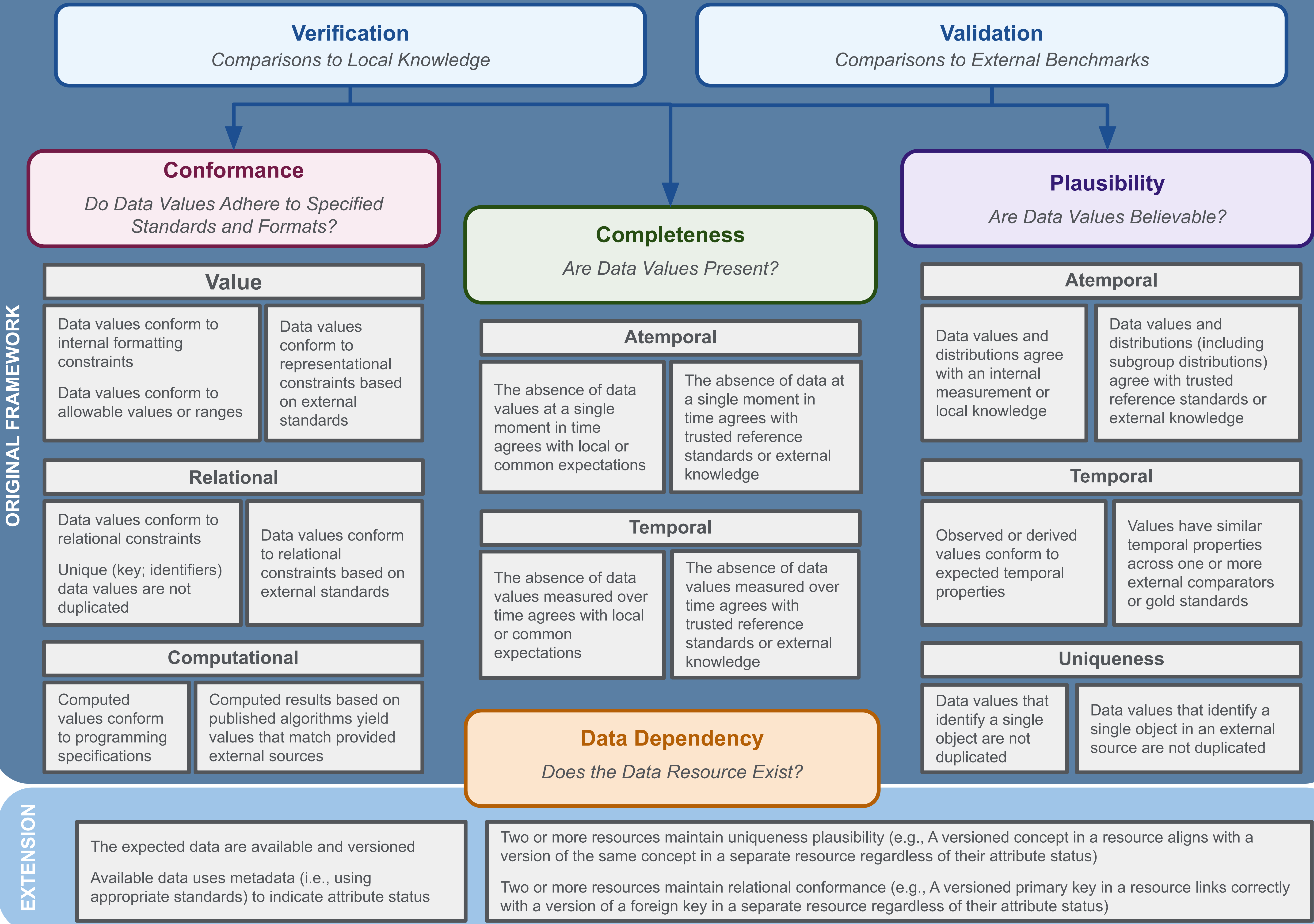
- Ontology checks were developed using a Jupyter Notebook and were reviewed by a member of the Open Biomedical Ontology (OBO) Foundry Working Group and a Semantic Web developer with >10 years of experience (11/2020-03/2021).
- Ontology checks were applied to 11 OBO ontologies: Chemical Entities of Biological 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).

### Ontology Check Characterization

- The original HDQF<sup>7</sup> was used for the ontology check mapping.
- Ontology checks were mapped using two contexts: (i) a single ontology and (ii) a set of merged ontologies.
- HDQF developers independently mapped the ontology checks.
- Any check not able to be clearly mapped was discussed until a final mapping consensus was reached.

Code, Notebooks, and data are available: <https://zenodo.org/record/6468948>.

## The Harmonized Data Quality Framework



## RESULTS

Table 1. Ontology Data Quality Checks.

Check	HDQF Alignment	Check Description
Deprecated Entities	<sup>a</sup> Data Dependency <sup>b</sup> 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	<sup>a</sup> Data Dependency <sup>b</sup> Relational Conformance	All deprecated and obsolete entities and associated triples are removed. For each ontology, removed entities are organized by status (i.e., obsolete vs. deprecated) and entity type (i.e., owl:Class vs. owl:ObjectProperty) in the ontology quality report. Issue: <a href="https://github.com/CLO-ontology/CLO/issues/48">https://github.com/CLO-ontology/CLO/issues/48</a>
Value Errors	<sup>a</sup> Value Conformance <sup>b</sup> 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 16". The erroneously defined literal was re-typed. Issue: <a href="https://github.com/PROconsortium/ProteinOntology/issues/176">https://github.com/PROconsortium/ProteinOntology/issues/176</a>
Semantic Heterogeneity	<sup>a</sup> Plausibility Uniqueness <sup>b</sup> 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 rdfs:SubClassOf 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	<sup>a</sup> Value Conformance <sup>b</sup> 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: <a href="https://github.com/PROconsortium/ProteinOntology/issues/176">https://github.com/PROconsortium/ProteinOntology/issues/176</a>
Identifier Errors	<sup>a</sup> Value Conformance <sup>b</sup> 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 HPO_ instead of HP_. 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 reference ontologies to match it). Erroneous identifiers are repaired via the use of a reference resource. In the VO, an incorrectly formatted occurrence of PRO_XXXXXX was replaced with PR_XXXXXX. Issue: <a href="https://github.com/vaccineontology/VO/issues/4">https://github.com/vaccineontology/VO/issues/4</a>
Punning Errors	<sup>a</sup> Plausibility Uniqueness <sup>b</sup> Relational Conformance	This check identifies entities with more than one type declaration. This check follows recommendations of Vrandecic (2010), <sup>9</sup> such that all entities are checked for redeclarations and when identified, the following corrections are made: I. Entities typed as an owl:Class and an owl:NamedIndividual are typed as an owl:Class II. Entities yped as an owl:ObjectProperty and an owl:AnnotationProperty are typed as an owl:ObjectProperty This check identified 7 illegally punning errors in the CLO. Issue: <a href="https://github.com/CLO-ontology/CLO/issues/43">https://github.com/CLO-ontology/CLO/issues/43</a>

<sup>a</sup>Individual ontology; <sup>b</sup>Merged ontologies. See ontology check code: [https://github.com/callahanitff/PheKnowl ator/blob/master/notebooks/Ontology\\_Cleaning.ipynb](https://github.com/callahanitff/PheKnowl ator/blob/master/notebooks/Ontology_Cleaning.ipynb).

## CONCLUSIONS

- 14 ontology checks were developed, evaluated, and mapped to the HDQF. The Data Dependency category was created to address dependency and versioning errors.
- 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.
- 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.

**Limitations:** (1) the ontology checks were tested on a small sample of ontologies; (2) existing ontology checks were not mapped to the HDQF. (3) the ontology quality pipeline does not include existing tools like ROBOT<sup>10</sup>; and (4) ontology developers did not actively participate in the HDQF alignment process.

## References

1. Blake et al. *J Biomed Inform.* 2006;3:314–20
2. Köhler et al. *BMC Bioinformatics.* 2006;7:212
3. Brank et al. *Conference on Data Mining and Data Warehouses.* 2005;166–70
4. Amith et al. *J Biomed Inform.* 2018;80:1–13
5. Parsia et al. *14th international WWW Conference.* 2005;633–40
6. Slater et al. *BMC Med Inform Decis Mak.* 2020;20(Suppl 10):311
7. Kahn et al. *eGEMs.* 2016;4:1244
8. Callahan et al. *eGEMs.* 2017;5:8
9. Vrandecic Z. *Karlsruher Institut für Technologie;* 2010
10. Jackson et al. *BMC Bioinformatics.* 2019;20:407

- The 14 ontology checks are shown in **Table 1**. Two iterations of revisions were required to finalize the checks.
- The CLO had the most errors and VO had the fewest. The amount of deprecated entities ranged across the ontologies with ChEBI (n=18,506), SO (n=341), and GO (n=6,430) containing the greatest proportion. Punning Errors were only identified in the CLO (n=16) and when merging the 11 ontologies (n=8).
- 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 check within the individual and merged ontology contexts are shown in **Table 1**. All of the checks except Deprecated and Obsolete Entities were mapped to an existing HDQF category which led to 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, 28.6% to Plausibility Uniqueness, and 28.6% to Data Dependency. Within the merged ontology context, all of the ontology checks mapped to Relational Conformance.