Response to Reviewers

We thank the reviewers for the fair and constructive comments on our original submission. We believe we have addressed their concerns, as spelled out in the point-by-point response below. The original reviewer comments are reproduced in *italics*, followed by our response.

Reviewer #1:

*This manuscript describes the Ontology for Biomedical Investigations or OBI, its rationale, its method of development, its adoption, and the how and the extent to which it is meeting its use cases. As the manuscript indicates, OBI has gained substantial momentum in recent years and is a highly valuable community resource. As such, dissemination of its rationale, development, use cases, adoption, community, and usage is of great merit and importance.*

*Methodology, Choice of metadata conventions: please add a description of what they are with examples. It is not clear here what the authors mean by 'metadata' conventions. Also, later in the manuscript, the authors might highlight the extent to which other ontologies have adopted them. If there is significant adoption of them, this could be highlighted as another contribution of OBI.*

We have added to the Methodology section to address the reviewer’s comments as well as made a more clear statement of the contribution of OBI in the Discussion. The choice of metadata conventions section now reads: "OBI developed and has consistently used a convention for information that documents classes (metadata). This takes the form of a number of documented annotation properties and fillers. Initially this metadata included basic information such as labels, synonyms and definition, attribution to editors and source for definitions. As OBI’s development progressed additional properties were added as need was recognized. Curation status gives an indication of the level of development of a term, and its possible values have been augmented as different cases emerged. Various forms of additional documentation properties have been added in response to needs of developers and users. These now include example of usage of a term, notes that aid the user in understanding the term, notes that related to development process. The metadata scheme is implemented as an OWL ontology available at [http://purl.obolibrary.org/obo/ontology-metadata.owl]. A specification of what is considered minimally acceptable use is available at [http://purl.obolibrary.org/obo/obi/policy/metadata]. This metadata scheme has now been adopted by a number of OBO ontologies and continues to evolve with participation from other members of the OBO community."

*line 184: reference 18 is incorrectly cited here. Reference 18 is a Foundational Model of Anatomy paper which seems completely irrelevant to NCBI Taxonomy. Perhaps the correct reference is #24?*

Thanks for pointing this out. We have updated the citation for #18. Additionally, reference 24 should have pointed to the same paper.

*Results, line 273: the definition of planned process gets a bit confusing here. Please clarify. Is it the case that only processes that meet one or more objective specifications of the plan specification fulfill the criteria? An appendectomy is a planned process only if the appendix comes out right? Not just if it is scheduled and someone intends to do it?*

*The reviewer is correct; we indeed only consider completed planned processes in which the objectives were fulfilled. We have added the example given by the reviewer to further clarify this. TODO*

8. Results, Organization, lines 341-350: no rationale is given for representing an organization as a material entity. Various debates are cited, but not why they were resolved in favor of material entity. Please add some rationale.

The reviewer is correct that rationale was missing. We have updated the manuscript to state: “OBI needs to represent organizations to identify, for example, Affymetrix as the manufacturer of a microarray. Placing organization into the BFO hierarchy proved controversial, as good arguments were made for treating it either as a material entity or an immaterial kind of social construct related to other legal entities. The latter are not currently well described in BFO. In fact, both exist, though in English the same word (organization) is fluidly used to describe one or the other. Something that is true is that any time the organization is comprised of a number of persons. That aspect can be captured as a material entity – the aggregate of those people (instances of homo sapiens, which are material entities). Organizations can also have other organizations as member and since those are also material entities, the whole is a material entity. Since material entities are better developed in BFO, we associated the label ‘organization’ with the material entity. The social aspects of the organization can be captured by organization-specific member roles, for example the role treasurer, or administrator, or compliance officer, although OBI has not found it necessary to define such roles. The current definition of organization is broad enough to include for example a regulatory agency like the U.S. Food and Drug Administration."

9. Results, line 359: what constitutes a "relevant journal article" for IEDB? What is the incentive for authors to contribute their data? Please elaborate a bit.

Response: The scope of the IEDB was established by the funding agency (NIAID) to encompass any publication relating immune epitope data where the epitope is relevant to allergy, infectious diseases, autoimmune diseases and transplant related illnesses. We have edited that section as follows:

The Immune Epitope Database (IEDB)[29] catalogs experiments that characterize the location and function of immune epitopes in infectious agents, allergens, transplants and auto-antigens. Information is entered into the IEDB through author submissions and through manual curation of the scientific literature. Authors submit data to the IEDB because they are contractually obligated by the NIH to do so or in some cases authors have requested submission of large datasets that are not very useful when published as unstructured supplemental material. Over 1,000,000 experiments have been entered into the IEDB to date and manual curation covers more than 98% of all journal articles containing epitope information, as described above.

10. Results, line 459: not every reader will be familiar with nanopublications, so either add a bit to explain or leave them out.

We added a citation to nanopublications to inform readers about this. [TAKE FROM EMAIL]

11. Discussion, lines 554-555 (and maybe 556): What is the rationale for reviewing the assay terms next? What is motivating it? What use case / user community / data / etc.?

*We have now added the rationale: “The assay terms were chosen as the next focus items as over the past years a large number of descendent terms have accumulated, and there is a need to review the different design patterns especially for the logical definitions to achieve better consistency.”*

12. Figures, in general the resolution was not great and I had a hard time reading text in them. Might be a conversion to PDF issue

ASSIGN TO EACH.

13. Abstract, line 48: suggest making the word 'meaning' plural as not all terms have a singular meaning

Response: We agree and have made this edit.

14. Abstract, line 48: the term 'biomedical investigations' is a bit vague. As the manuscript indicates, the authors use it quite broadly. This should be made clear at some point in the manuscript, and as much clarification as can be made to fit in the abstract would be helpful.

We have updated the abstract to explicitly state: “The Ontology for Biomedical Investigations (OBI) is an ontology that provides terms with precisely defined meanings to describe all aspects of how investigations in the biological and medical domains are conducted”

15. Introduction, lines 107-112: use consistent capitalization for 'OBO Foundry'. It's a proper noun, so both words ought to be capitalized.

Response: We corrected all instances to be 'OBO Foundry.'

16. Methodology, line 125: 'homogeneous' is misspelled (missing the second 'e'). Please fix.

Response: This edit has been made.

17. Methodology, line 222: the grammar is not correct here. Perhaps "We are in the process of implementing checks that verify that all IDs that were present..."

Response: We have reworded the section as follows:

We are implementing checks to verify that all IDs present in the previous release are still in use as part of the release process.

AR: Yes, but we've already implemented this and we used it for a number of releases when I was helping do them. The current work would be a second implementation.

18. Results, figure 1, lines 237-240: the caption references blue, dark red, orange, and purple boxes, but only blue and dark red boxes are present. Either the figure is incomplete or its description is overly inclusive of elements not present in the figure.

*ASSIGN FIGURE*

19. Results, table 1: What is the rationale for the values of curation status annotation? How were these values chosen and how are they used? An example would be nice.

We have added the following text to the “Choice of metadata conventions” section in order to provide a rationale for the choice of Curation Status values: “The metadata scheme specification includes definitions of the curation status values (shown in Table 1). Their choice is reflective of the discussion about the term occurring amongst OBI developers. For example, terms that have been discussed on an OBI developers call and agreed upon for inclusion are ‘ready for release’. Terms that have been added following an agreed upon pattern but not reviewed by anyone but the term editor are ‘pending final vetting’. “

20. Results, line 397: would say "a NIAID Bioinformatics Resource Center" not "an NIAID..."

Response: "an" goes before a word beginning with a vowel sound. The \*letter\* "n" sounds like "e-n" when just speaking the letter. http://www.englishpage.com/articles/a-vs-an.htm

21. Results, lines 418-19: grammatical problems. Please fix. Specifically, "...capturing phenotype information resulting experiments in which..." is awkward and confusing.

Response: We have edited the section as follows:

In OBI, we describe the genotype of an isolate by first referring to the process of specimen collection that resulted in the physical isolate, followed by the sequencing experiment that was performed on the isolate, which provided the genotype information.

And

The same approach is used for the more complex task of capturing phenotype information obtained by genetically modifying parasites.

22. Results, figure 4, lines 427-433: this is a diagram of instances, and not types, correct? Not every assay as a "product of gene of interest" as specified input, after all. Either way, would clarify.

*The reviewer is correct, this is a diagram of instances; the label in the boxes indicates a class that is being instantiated. The Figure 4 legend has been updated accordingly to include: “The representation is at the instance levels, i.e. not all assays will have the specified inputs and outputs. The class mentions are to indicate what is being instantiated.”*

*AR Will send comments*

23. Results, line 445: change word 'representation' to plural

Response: This was done.

25. Results, lines 455-457: gets confusing at points. Suggest editing for clarity. You mean the developers of ISA requested that terms be added to OBI for measurement and technology types, correct? What is an annotation in ISA vs. an annotation in an OWL ontology? How specifically is the ISA 'protocol type" supported by OBI terms, and which ones?

Response: The section was edited as follows: not yet edited

The format is domain agnostic, but needs controlled terminology for ISA ‘measurement’ and ‘technology types’ when defining assay tables. Term submission from ISA to OBI addressed those needs. Similarly, ISA ‘protocol type’ annotation is supported by specific OBI terms. The default ISA configurations (templates for creating ISA-Tab files) provided with the tools rely on OBI terms (Fig 5 panel a and b).

*PHILIPE please edit abov*24. Results, lines 446, 449: fix grammar.

Response: The section was edited to: not yet edited in manuscript

The ISA model is supported by a set of software applications to configure annotation requirements and create ISA tables, import information into databases and convert ISA representations into a range of formats for deposition into public repositories (e.g. ENA [36], ArrayExpress [37]). The workflow of the ISA system has been validated and is compatible with existing technology centric formats (e.g. MAGE-TAB for ArrayExpress [37] and SRA-XML for ENA [36], respectively). It can be applied to import data into analysis environments or publish the data alongside a narrative article [38-41] .

*PHILIPE please check for accuracy; this was butchered to pull it into multiple sentences for readability.*

*e*

*AR: Should give parent terms below which terms can be taken, for each field. E.g. if ISA's measurement can be filled by any assay, say so.*

26. Results, line 464: is 'curatorial' a word?

Response: Curatorial is a word. You can get a degree “Curatorial Practice”<https://www.cca.edu/academics/graduate/curatorial-practice> and there are many other references:<https://en.wikipedia.org/wiki/Curatorial_platform>

27. Results, lines 471-473: grammatical issues.

Response: The section was edited as follows: not edited

While the generic RDF serialization engine ensures conversion of ISA assays based on standard patterns, refinements and specialization can be added by supplying a new mapping between domain specific ISA configurations and OBI representations, thus granting more precision, as OBI and ISA continue to develop. An example of ISA extensions to targeted metagenomics application and mapping to OBI and OBO Foundry resources has been discussed in a detail in [43].

*PHILIPE please edit above*

*AR: Generic RDF serialization ensures conversion of ISA assays based on standard patterns. Refinements and specialization can be added by supplying a new mapping between domain specific ISA configurations and OBI representations, thus granting more precision. This extensibility means that representations can improve as OBI and ISA continue to develop. An example of ISA extensions to targeted metagenomics application and mapping to OBI and OBO Foundry resources has been discussed in a detail in [43].*

28. Discussion, line 505: suggest changing 'between' to 'among' as there are surely more than two individuals.

Response: This edit has been made.

29. Discussion, line 545: suggest changing 'these' at end of sentence to 'them'.

Response: The line has been edited as follows:

The tools can set the minimal fields[3] required and restrict each of them to appropriate terms from OBI or other OBO ontologies without requiring the end user to have an in-depth understanding.

30. Discussion, lines 563-4: suggest setting off the clause "which has been recently reviewed" in commas to improve readability.

Response: The edit has been made.

31. Also who reviewed the core of OBI?

*The review was conducted internally. This was now added to the manuscript:*

*However, the core of OBI, which has been recently reviewed internally, by the OBI team, is considered stable with regards to the asserted hierarchy.*

32. Acknowledgements, line 576: OBO Foundry capitalization issue again.

Response: The edit has been made.

33. Acknowledgements, line 578: add period at end of last sentence.

Response: The edit has been made.

Reviewer #2:

34. This manuscript describes the Ontology for Biomedical Investigations, and represents the efforts of a broad community to establish a precise semantics for describing biomedical investigations. The manuscript is clearly written, and several use cases of the ontology are presented. The manuscript is relevant to current discussions about data sharing and interoperability in the broader scientific community.

*Response: some thanks go here*

35. MIREOT is mentioned as a solution to importing large external ontologies. This appears on p 5 and again in the Discussion, however, it is not adequately explained in either place. What does it do? How is it used?

*We have edited this section, added more detail on the technical aspects of the MIREOT mechanism, and moved the section about deprecation policy which is not integral to that mechanism.*

*The section now reads as follows:*

*Importing such ontologies as a whole can lead to inconsistencies or unintended inferences. Instead, the MIREOT import mechanism was developed[19], which relies on a deprecation policy following the practice of the Gene Ontology in that the intended meanings of classes remain stable. Even when the source ontology changes, such modifications are not intended to change the meaning of terms. We use the OntoFox [20] tool to facilitate the import of 433 terms from external ontologies into OBI with the MIREOT technique.*

*AR: OBI has also spawned other ontologies (Information Artifact Ontology) as well as methods for importing only designated terms from external ontologies (MIREOT).*

*Importing such ontologies as a whole can lead to inconsistencies or unintended inferences. Instead, the MIREOT import mechanism was developed[19]. MIREOT allows specification of a set of terms to be imported including mapping of metadata, placement of imported terms within the OBI hierarchy and inclusion of selected axioms. We use the OntoFox [20], which acts on these specifications, to facilitate the import of 433 terms from external ontologies into OBI with the MIREOT technique. OntoFox is re-used for each release ensuring that imported terms are kept up-to-date with their source.*

*<p>*

*Our deprecation policy follows that of the Gene Ontology in that the intended meanings of classes are to remain stable. Even when the source ontology changes, such modifications are not intended to change the meaning of terms. Ontologies from which we import terms are reviewed to ensure this is a valid assumption.*

36. in general the section on p 5 that discusses "Integration with existing ontologies" is a little shallow; at this point in the manuscript the idea of/need for incorporating external ontologies has not been introduced -- it is simply assumed "... useful when integrating external ontologies." What external ontologies might be relevant? Why would you need to import an ontology when it can be referenced via IRIs?

Revisions by James Overton:

Integration with existing ontologies

OBI was developed to be complementary to, and integrated with, a framework of existing ontologies in the biomedical domain. For example, when an investigation involves a subject organism we include the NCBI Taxonomy term for its species or strain, and when an assay measures a biological process we include the Gene Ontology term for that process. Whenever possible, we reuse existing terms rather than creating our own, allowing our users to take advantage of data and annotations that already use those existing terms. This integration is more than simple linking: we incorporate external ontology terms into our definitions, logical axioms, and annotations, to create a larger framework that facilitates data integration and reuse across biology and biomedicine. The methodological challenges associated with this reuse are discussed here.

A top level ontology can provide guidance in how individual terms or whole ontologies interrelate and is useful when integrating external ontologies. BFO[16] was chosen as the top level ontology as it is stable, publicly available in OWL syntax, and aligned to existing OBO Foundry ontologies, and because there is a community of developers and users. The OBO Relations Ontology[17] provides relations for OBI. When new relations have been defined between classes in OBI, these are based on RO relations wherever possible. The Information Artifact Ontology developed out of OBI and the two ontologies continue to have a close relationship. OBI uses the OWL import mechanism to import BFO classes, RO core relations, and all of IAO, thereby integrating all those terms, their axioms, and their annotations into OBI.

The OWL import mechanism is appropriate when importing all terms from the target ontology. It not suitable for all imports of external ontologies, for two reasons: first, current editing tools are not effective for working with very large ontologies such as the NCBI Taxonomy[18], therefore a direct import is not scalable; second, some ontologies used by OBI are actively developed and may not be aligned with OBI methodology, for example they may not use BFO or OWL DL. Importing such ontologies as a whole can lead to inconsistencies or unintended inferences. Instead, the MIREOT import mechanism was developed[19], which relies on a deprecation policy following the practice of the Gene Ontology in that the intended meanings of classes remain stable. Even when the source ontology changes, such modifications are not intended to change the meaning of terms. We use the OntoFox [20] tool to facilitate the import of 433 terms from external ontologies into OBI with the MIREOT technique.

37. Policies (metadata and deprecation, etc.) are mentioned in the "Quality control" section; please summarize them as the reader is likely to appreciate what the quality issues are.

*ASSIGN*

38. the discussions of the use cases may need some reworking of the tenses -- e.g. the work of managing vocabulary for the IEDB is described in the present tense as a thankless task, and implies that the IEDB is (currently) not useful

*ASSIGN*

39. outside of the context of the database itself. It left me with the question "So why bother?" but then I realised the text refers to the non-/pre-OBI representation in the database. I think reframing to indicate that the OBI enabled a change from something that as only internally useful to something that was easier to maintain and interoperable with other resources would be an improvement. Same issue occurs in the description of the EuPathDB ("is hard to query" cf "was hard to query").

*ASSIGN*

40. The mention of "logical reasoning organizing terms into a hierarchy" could be elaborated (p 10, line 379) -- how does this happen?

*ASSIGN*

41. in general, I'd like to see more discussion of the broader impacts of adopting the ontology for the highlighted databases -- impacts for re-use, reasoning, etc. This is implied in several places ("hard to query") but not a specific focus of the use case discussions -- has re-use of the data/databases in other contexts been facilitated in practice?

*ASSIGN*

42. p 13 lines 525-539 are replete with acronyms that are not expanded; it would be nice to expand them/say a little more here. Not all readers will be familiar with them or their relevance.

Response: We agree and have explained these and several other acronyms used throughout. Consider adding abbreviations list?

43. Typos/formatting: p 9 line 360 refers to the year 2011 -- this seems incorrect in the context where it is used. Or is it complete?

Response: This section has been revised as follows:

The Immune Epitope Database (IEDB)[29] catalogs experiments that characterize the location and function of immune epitopes in infectious agents, allergens, transplants and auto-antigens. Information is entered into the IEDB through author submissions and through manual curation of the scientific literature. Authors submit data to the IEDB because they are contractually obligated by the NIH to do so or in some cases authors have requested submission of large datasets that are not very useful when published as unstructured supplemental material. Over 1,000,000 experiments have been entered into the IEDB to date and manual curation covers more than 98% of all journal articles containing epitope information, as described above.

44. Citation formatting is inconsistent: some have spaces before them and some don't. Please review.

*Response: The formatting is controlled by aliens?*

45. There is one reference to a citation/PMID on p 14 line 447 that is unspecified.

*Response: I cannot tell what they mean. Is now line 463*

*ASSUME PMID25052702*

46. p 6 line 209: "workshop" -> "workshops"

Response: This edit has been made.

47. p 6 line 222: "check"->"checks"

Response: This edit has been made.

48. p 6 line 224: "follows GO" -> "follows the GO"

Response: This edit has been made.

49. p 13 line 550: "as science" -> "as in science"

Response: This edit has been made.

50. p 13 line 517 "craft knowledge" -> "expert knowledge" ?

Response: This edit has been made.