NGBO: The Introduction of -omics Data to Biobanking Ontology

Dalia Alghamdi[1, 4], Damion M. Dooley [1], Gurinder Gosal[1], Emma J. Griffiths[2], William W.L. Hsiao[1-3]

[1] University of British Columbia, Vancouver BC V6T 1Z4, Canada.
[2] Simon Fraser University, Burnaby, BC V5A 1S6, Canada.
[3] BC Centre for Disease Control Public Health Laboratory, Vancouver, BC V5Z 4R4, Canada.
[4] King Fahad Medical City, Riyadh 59046, Saudi Arabia.

Dalia.Alghamdi@bccdc.ca

Abstract. A biobank contains a collection of biological samples, along with associated medical information of sample donors, which can be used for different types of studies. Given the wealth of information that can be derived from stored information and biological materials, there is a pressing need for structuring biobank data for more computer-amenable analyses. The utility of first generation biobanks was originally evaluated simply based on the number of samples that they contained. Currently, the value of biobank data lies in how it can linked with other molecular and clinical data ("-omics data"), to provide new insights into health and disease. Linking data has thus far, however, proven challenging due to unstructured and incompatible data types. Here, we describe the development of a Next-Generation biobanking ontology (NGBO) (https://github.com/Dalalghamdi/NGBO) that is capable of supporting both Biospecimen processing, management, storage and retrieval infrastructure, and acting as a knowledge hub for an integrated clinical and translational research ecosystem integrating –omics data. NGBO harmonizes the instrumentation and procedures used to prepare and process specimens, and also covers terminology used to describe computational biology algorithms, analytical tools, electronic-communication protocols, in vitro assays. Laboratories, investigators, and other biobanks would also benefit from the knowledge contained in the ontology, by the means of using NGBO a biobank data catalogue that can be used to map any existing unstructured data.

Keywords: Ontology, Biobank, Next generation sequencing, data harmonization, data integration.

A biobank consists of various biological samples linked to the medical information of sample donors which can be used for translational and biomedical research. Biological samples can include organs, tissues, cells and body fluids [1]. The stored specimens enable researchers to save time and resources of collecting and processing new samples needed in there projects, thereby, they improve research outcomes, promising for more effective diagnosis and treatments of patients who suffer from common or rare diseases [2]. The collection of human tissues for the purpose of biomedical research began centuries ago, but has since undergone dramatic change due to technology advancements in storage techniques, sample information retrieval, as well analysis of specimen material. As such, the informatics needs of modern biobanks are far more complex than past repositories that often captured only the date or location of a sample collection. Besides sample metadata, modern biobanks cover the storage and management of more complex data generated from high-throughput biological studies such as proteomic, genomics and other –omics studies [3]. International and national collaborations can improve the value of biobanks, but this requires harmonizing the data fields and values across biobanking applications. There have been several efforts to achieve collaboration and data sharing among various national biobanks, for instance, the public population project in genomics (P3G) (http://www.p3g.org) has previously tackled building biobanking resources as well as data cataloguing and harmonization for data integration [4]. Still, biobanks remain heterogeneous when it comes to their design, usage, size and types of the samples. It is possible to link the samples to data records from expansive epidemiological collections and family histories. However, it is laborious to manually harmonize the terms across different biobanks. Furthermore, if data harmonization is conducted individually, inconsistencies often arise. A key development in facilitating data standardization is the application of ontology, a semantic web technology [5].

Semantic Web is the best practices and sets of standards used to share data and meanings (semantics) of data over the web. The formal and machine-readable definitions and axioms made it is possible to come up with automated querying systems to facilitate faster, easier, and more accurate ways to share and reuse data [6]. Semantic Web OWL ontology is a popular technology choice for representing terminology and data structure relationships. If one is to use ontology, it becomes possible to establish vocabularies necessary to model a problem or activity domain. In the model, there are objects and concepts contained in specific areas and relations describing how they are related [7]. Ontologies play a role in promoting the realization of Semantic Web.

Brochhausen et al. proposed the ontology for biobanking (OBIB) in 2016. The OBIB was created through the merging of two biobank technologies including the Biobank Ontology (BO) and the Ontologized Minimum Information About Biobank Data Sharing (OMIABIS) [8]. BO and OMIABIS focus on specimen description and biobanking administration respectively. The biomedical and biological research has progressed to a level where the quantity and the types of samples kept are no longer used in measuring the prowess of biobanks. Instead, measurements are based on the extent that the samples and metadata are used. Biobanks fall into the category of infrastructure used in research. Thus, their aim is to support scientific processes.

The integration of more knowledge domains into the biobanking ontology is necessary for advancing the use of biobanks. We aim to integrate –omics data, with the creation of Next-Generation Biobanking Ontology (NGBO) to deal with various scientific research and personalized medicine requirements. To incorporate -omics knowledge, we model the processes and entities related to diagnostic molecular pathology procedures, including sample handling, phenotype characterization, computational biology algorithms and analytical tools, in vitro assays, electronic communication protocols and data coding. In addition to the technical data provenance, sample data provenance such as patient phenotypic information, along with the genetic data will provide the biobank users sufficient data and knowledge to characterize the functional and pathogenic significance of genetic variants [9].

NGBO is being built based on the Open Biological and Biomedical Ontologies (OBO) foundry principles. For example, one of the OBO foundry principles is the re-use of existing ontology to prevent re-inventing the wheel and creating multiple representations of the same term. OWL (W3C Web Ontology Language) is used to provide the means for data sharing and reusing between different resources in the form of semantic application. NGBO will provide standard identifiers for classes and relations within biobanking domain as well as definitions for all the vocabularies in NGBO in human and machine-readable formats. Consider the class 'input data' as an example, the definition of input data that can be read by human is "computer file that has specific format and contains data that serve as input to a device or program". However, it could be defined using OWL language as:

```
'is about' some entity and 'has format' exactly 1'file format'
```

This is one of the expressions possible for the class 'input data', which states that it is about an entity and has only one file format from file format subclasses. As shown in figure 1, NGBO re-use many terms from existing ontologies such as Bio-Assay Ontology (BAO) and the OBO edition of the National Cancer Institute Thesaurus (NCIT) ontology. NGBO depends primarily on the (is-a) relation between classes and subclasses, thereby providing a hierarchy of classes that also enables inheritance of the properties. For example, a concept "planned process" (OBI:0000011) in the Ontology for Biomedical Investigations (OBI) is defined as "a processual entity that realizes a plan which is the concretization of a plan specification". Therefore, all sub-classes of planned process must inherit the definition of the process. In addition to (is-a), pre-existing relations such as (is_version_of) is used when needed. Protégé (version 5.2.0) is used to build NGBO that is compatible with the Basic Formal Ontology (BFO), a small upper level ontology mostly used to support information retrieval, analysis and integration in biomedical and biological domains [10]. Figure 1 shows an example of selection of NGBO classes and their sub-classes.

In conclusion, by building NGBO, a next generation biobanking ontology, we are providing a semantics infrastructure to support externalized biomedical collaborative research by harmonizing biospecimens with their molecular makeup. It provides a framework for reusing clear consistent terminology (classes) with their relationships and the metadata that describe the intended meaning of these classes and relationships.

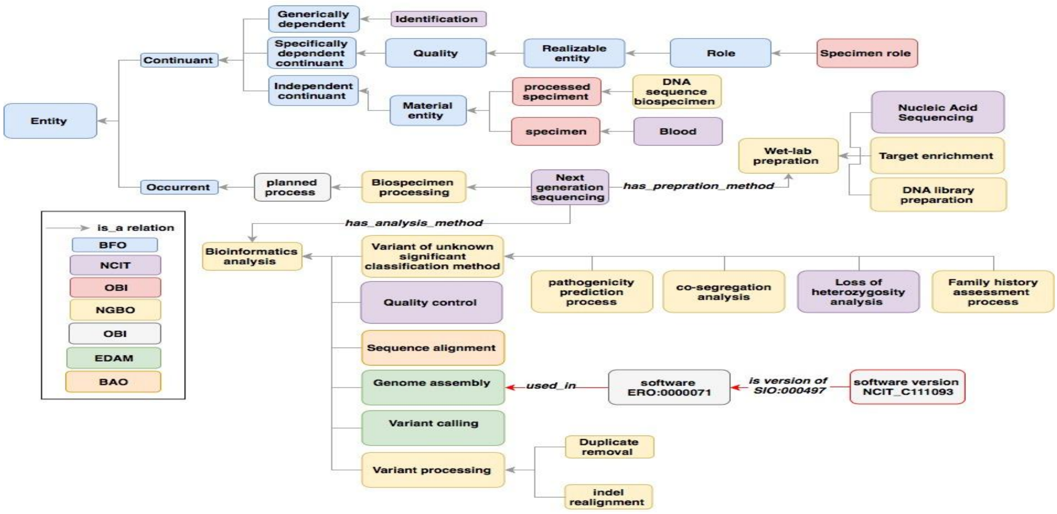

Figure 1: The selection of main NGBO classes and their sub-classes. For readability reasons, the leftmost classes are missing in the figure.

References:

1. Elger, Bernice S., and Arthur L. Caplan. 2006. “Consent and Anonymization in Research Involving Biobanks: Differing Terms and Norms Present Serious Barriers to an International Framework.” *EMBO Reports* 7 (7): 661–66.
2. Riegman, Peter H. J., Manuel M. Morente, Fay Betsou, Pasquale de Blasio, Peter Geary, and Marble Arch International Working Group on Biobanking for Biomedical Research. 2008. “Biobanking for Better Healthcare.” *Molecular Oncology* 2 (3): 213–22.
3. De Souza, Yvonne G., and John S. Greenspan. 2013. “Biobanking Past, Present and Future: Responsibilities and Benefits.” *AIDS* 27 (3): 303–12.
4. Ouellette, Sylvie, and Anne Marie Tassé. 2014. “P(3)G - 10 Years of Toolbuilding: From the Population Biobank to the Clinic.” *Applied & Translational Genomics* 3 (2): 36–40.
5. Andrade, André Q., Markus Kreuzthaler, Janna Hastings, Maria Krestyaninova, and Stefan Schulz. 2012. “Requirements for Semantic Biobanks.” *Studies in Health Technology and Informatics* 180: 569–73.
6. Tim Berners-Lee, et al, “The World Wide Web,”Communications of the ACM, August, 1994.
7. T. Gruber. 1993 “A translation approach to portable ontology specification”. “Knowledge Acquisition”, pp.199-220.
8. Brochhausen, Mathias, Jie Zheng, David Birtwell, Heather Williams, Anna Maria Masci, Helena Judge Ellis, and Christian J. Stoeckert Jr. 2016. “OBIB-a Novel Ontology for Biobanking.” *Journal of Biomedical Semantics* 7 (May): 23.
9. Johnston JJ, Biesecker LG. Databases of genomic variation and phenotypes: existing resources and future needs. Human Molecular Genetics. 2013;22(R1):R27-R31. doi:10.1093/hmg/ddt384.
10. Arp, Robert, Barry Smith, and Andrew D. Spear. 2015. “Building Ontologies with Basic Formal Ontology.” The MIT Press. https://dl.acm.org/citation.cfm?id=2846229.