Quality of Life through Quality of Information
J. Mantas et al. (Eds.)
IOS Press, 2012

doi:10.3233/978-1-61499-101-4-569

Requirements for Semantic Biobanks

André Q ANDRADE[a,b,1], Markus KREUZTHALER[b], Janna HASTINGS[d,e], Maria KRESTYANINOVA[f,g] and Stefan SCHULZ [b,c]
[a]*School of Information Science, Federal University of Minas Gerais, Brazil*
[b]*Medical University of Graz, Austria,* [c]*University Medical Center Freiburg, Germany*
[d]*European Bioinformatics Institute, Hinxton, UK;* [e]*University of Geneva, Switzerland*
[f]*Helsinki University, Finland,* [g]*Uniquer, Lausanne, Switzerland*

Abstract. World-wide availability of biobank samples is a great desideratum for biomedical researchers. We describe the use case of biobank information retrieval that requires the semantic descriptions of biobank samples and of clinical information. In addition we sketch the foundations of an ontology for biobanks, as a basis on which distributed biobank indexing and retrieval systems can be built. We advocate that a detailed and robust representation of this kind of information improves and allows complex queries that will certainly arise to explore the full potential of biobanks.

Keywords. Ontology, Biobanks, Description Logics, Information Retrieval

Introduction

Whereas, historically, the sampling and analysing of body material has been primarily relevant for diagnosis, biosamples (e.g. blood, tissue) are becoming increasingly important for research. This has stimulated the development of biobanks, systematic collections of tissue samples and population-wide data on health and lifestyle. Biobank management systems provide search functionality to facilitate access to sample data and metadata. To obtain statistical effectiveness for a particular research question, it is often necessary to use samples and data from more than one biobank. This is a difficult challenge not only due to semantic heterogeneity across different sites, but also due to differing regulatory contexts and languages. Furthermore, Biobanks are overly heterogeneous regarding research targets, sorts of samples, and available data from different populations. Searching for relevant samples across different biobanks is currently a laborious process. Even if a technical solution is put in place, its usability will fundamentally depend on the completeness and the trustworthiness of the underlying metadata annotations. Advanced informatics methods are therefore vital for handling samples and their related metadata.

A key requirement is to agree upon a common semantic foundation for storing and communicating sample description metadata. This creates a foundation for harmonization activities undertaken by various initiatives like P3G [1] and by stand-alone biobanks. In this paper we will investigate how a generic structure based on

[1] Corresponding Author. André Q Andrade, School of Information Science, Federal University of Minas Gerais, Brazil

formal ontologies can support the process of harmonising the semantic content of biobanks.

1. Methods

The biomedical information explosion has prompted the development of ontologies, visible by the continuously growing BioPortal library, as well as by the OBO Foundry. Ontological methods are also increasingly using terminological standards like SNOMED CT and WHO-FIC classifications. Description logics (DLs) [2], often using OWL [3], have become a quasi standard for formal ontologies, which intend to describe (as much as possible) the consensus on the nature of entities in a given scientific domain, independently of linguistic or conceptual variation. Examples of statements belonging to this consensus core are indisputable truths like: all cells contain membranes, all metastases derive from some primary tumour, and all sampling events have some sample as their outcome. Ontology construction should obey principled criteria, enforced by top-level ontologies. We are using BioTop [4] as the basis of our attempt to represent central notions of sampling in the context of biobanks.

All samples have in common that they are material objects which derive from some parts of an organism. However, the sample may survive the organism. Through its processing (e.g. freezing), a sample is considered to become a new entity when it is stored in the biobank. Each sample is the outcome of some biosampling event.

To evaluate the required domains to represent, store and query biobanks, we considered the current requirements and practices described in [5] and created a series of queries to test possible representations. Domains were identified using a generic ontological approach, distinguishing entities in the reality (clinic, lab) from information and epistemological entities (entries in documentation systems).

In order to make an objective evaluation of the proposal, we will first put forward a clear use case for indexing biobanks samples and for addressing related retrieval scenarios. Biobanks will mostly be used to store slow decaying or static samples of different organisms' tissues, which are subsequently retrieved to be subject to specific tests that had not been performed at the time of sampling, e.g. due to the development of new diagnostic techniques. There are two main types of queries: clinically oriented and pathology oriented ones. The latter are geared toward sampling techniques, sample preparation and storage, which can meet several grades of quality. The former inquire about all health-related events of the patient whose sample is stored in the biobank. They are closely related to long-term patient follow up and should be continuously fed by electronic health records maintained by clinicians. Often, queries will be a combination of both. We propose for evaluation the following queries:

- retrieve all frozen gastric mucosa samples of patients who had cancer of stomach after 2008;
- retrieve all HE stained biopsy sample from the antrum mucosa;
- retrieve all hepatic samples from patients with stomach cancer before 2009, confirmed by biopsy, with no death event is registered before end 2011

Another required additional aspect (metadata) regards information about access control of samples. While normally embedded into the system, we envision that important information indirectly related to samples and patients must be ontologically represented. Privacy requests should be associated with the samples as limiting factors for full

access, an issue addressed, e.g. by the Human Sample Exchange Regulation Navigator [6]. An interface between online resources that facilitate legal and ethical regulations and those that represent other metadata related to samples is yet to be established.

We should distinguish between information artefacts and entities in reality, as proposed by IAO [7]. That means to make explicit the distinction between information such as "sample s001 is stored in a freezer" and "a record entry on s001 was created by user001". That kind of distinction is particularly useful for assigning confidence values to information, which has to be addressed by the search mechanism. For instance, manually coded data are more reliable than data automatically extracted from narratives. Sample related information is given by attributes that describe the actual collection, processing and storage process. A clear identification of standard procedures of biobanks (e.g. freezing, formalin fixed paraffin embedded) and common sample origins (tissues, blood and derivatives) is therefore necessary to support the retrieval process. The annotation of anatomical structures and body fluids can be supported by several ontologies, such as the Foundational Model of Anatomy [8] and the "specimen" and "body structure" hierarchies of SNOMED-CT [9].

Finally, the clinical domain refers to all clinical information relevant to the sample donor. Such information can be related to the current disease that led to the sample collection, but also refer to past history, medication, family history and outcome after sample collection. Such a broad scope requires high-coverage and expressive ontologies that can convert complex statements to a commonly agreed meaning.

We provide further examples of the sample and clinical domain on the discussion section, using the sample queries as testing ground.

2. Results

Based on OBI, the Ontology of Biomedical Investigations [10] we can create two concurrent descriptions of the same event:

BioSampling equivalentTo *MaterialSamplingProcess* and
(**hasSpecifiedInput** some *MaterialEntity*) and
(**hasPatient** some *BiologicalStructure*) and
(**hasSpecifiedOutput** some *BiobankSample*) (1)

BioSampling equivalentTo *Action* and
hasPatient some *BiologicalStructure* and
hasOutcome some *BiologicalSample* (2)

In the first one, we use the OBI relations which relate input and output with a process, according to some objective specification. However, the relation **hasOutcome** is defined by BioTop as "participant which either - a) comes into being during the process or - b) undergoes some change during the process and constitutes (one of) the main result(s) of the process". As explained above, it is advantageous to represent the biobank sample as something new, created in the process of sampling. Since the specified input relation requires that there is no creation of the participant, we chose the BioTop relation. We also use the BioTop relation **hasPatient**, defined as a relation between "a participant (and) a process, with the condition that that this participant is not causally active". Since it requires determining whether the participant is causally active, this relation is sometimes difficult to identify. However, in the biobank case, there is a fundamental distinction between agents in the process (healthcare

professionals processing samples) and patients in the process (samples being processed), which is useful for correctly representing the relation and querying the information afterwards.

There are special sampling techniques, e.g. the taking of a biopsy, for instance[2]:

BiopsySampling equivalentTo *BioSampling* and **hasOutcome** some *BiopticSample* (3)

Other action classes are the preparation of a sample, e.g. staining, and subclasses such as *SampleStaining* (subclass of *SampleProcessing*) and *SampleHEStaining* (subclass of *SampleStaining*). Biological samples are the output of some sampling events in which some biological material which is part of an organism participates. We also introduce the OBO relation **derivesFrom**, which is implied by the concatenation of **outputOf** and **hasParticipant**. So we can express, a statement such as "HE stained biopsy sample from the antrum mucosa" with the following logical formulation[3]:

Sample and **patientOf** some *SampleHEStaining* and
derivesFrom some *AntrumMucosa* (4)

Due to the above axioms this is also retrieved by a query for a sample that derives from some gastric mucosa. In this way, we can achieve the expressivity required by common queries, which are mostly concerned about the original location/organ of the sample, while still maintaining the ontological definition that samples are not part of someone. Finally, the query mentioned above, *viz.* "retrieve all frozen gastric mucosa samples of patients who had cancer of stomach after 2008" would then look like[4]:

Sample and **participantOf** some *FreezingProcess* and
outputOf some (*Sampling* and
startTime some datetime [<2003] and
hasParticipant some (*MaterialBiologicalEntity* and
properPartOf some (*GastricMucosa* and
properPartOf some (*Human* and
participatesIn some (*HumanLife* and
hasProcessPart some *(IntestinalCancer* and
startTime some datetime [>2008])))))) (5)

This query would then e.g. retrieve a frozen sample collected in 2002, obtained from the antrum mucosa of a patient who eight years later got a polypoid adenocarcinoma of the duodenum.

Once the query is formulated, we require a mechanism for converting the description logic assertion into specific database queries of each institution. The biobanking network is likely a federated system, and most information will be stored locally. A similar approach for biobank data and metadata collection and for resource querying was implemented in [11]. SAIL, Sample avAILability system which initially was implemented in the ENGAGE project [12], and subsequently used for case studies in ELIXIR [13] and BBMRI [5], aimed to collect structure information about sample availability across European population biobanks. There was no ontological structure underneath, except a very generic domain language which required an expert curator between biobank data provider and the system. Standardised semantic content of SAIL was to be provided by DataSHaPER [14], EFO [15], OBI, and other initiatives aiming to create ontological representation of biomedical domain, the main mission of the platform was to demonstrate the added value of normalized phenotype descriptors and

[2] we avoid the term "biopsy" because it is used for both BiopsySampling and BiopticSample.
[3] **patientOf** is the inverse relation of **hasPatient**, referring no non-agentive participants of a process.
[4] The queries were formulated in OWL Manchester syntax.

the need for traceability of various harmonization efforts undertaken by sample collections. In addition to standard data schemas and ontologies the SAIL platform permitted data providers to establish their own, study-specific, harmonized vocabularies, i.e. candidates for ontologies. However, upon implementation and the first release of the system it was shown that only rigorous ontological formulation guarantees the common ground for information representation, allowing better data connectivity, query flexibility and reliable inferences on biobank data.

3. Conclusion

We have shown that semantic representation of biobank information is imperative for efficient integration of different domains and that there is a direct benefit to querying and searching of biobank databases. Although several harmonization platforms for biobanks have been released in the last years, none of them so far has benefited from clear and complete ontological formulation of semantic information. We developed a new, easy-to-implement and easy-to-follow, formalism that enables biobanks to enrich, re-annotate and make universally searchable the information about their contents and related clinical data.

Acknowledgements. AQA was being financed by CAPES (Brazil) – Programa de Doutorado no País com Estágio no Exterior, process number 2380-11-0, during the writing of this paper.

References

[1] Knoppers BM, Fortier I, Legault D, Burton P. The Public Population Project in Genomics (P3G): a proof of concept? Eur J Hum Genet. 2008; 16(6):664-5.
[2] Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF, editors. The Description Logic Handbook: Theory, Implementation and Applications. 2nd ed. New York: Cambridge University Press; 2007.
[3] Hitzler P, Krötzsch M, Parsia B, Patel-Schneider PF, Rudolph S. OWL 2 Web Ontology Language: Primer W3C2009 27 October.
[4] Beißwanger E, Schulz S, Stenzhorn H, Hahn U. BioTop: An Upper Domain Ontology for the Life Sciences - A Description of its Current Structure, Contents, and Interfaces to OBO Ontologies. Applied Ontology. 2008; 3(4):205-12.
[5] Yuille M, van Ommen GJ, Bréchot C, Cambon-Thomsen A, Dagher G, Landegren U, Litton JE, Pasterk M, Peltonen L, Taussig M, Wichmann HE, Zatloukal K. Biobanking for Europe. Brief Bioinform. 2008;9(1):14-24.
[6] http://www.hsern.eu/ (last accessed April 13th 2012)
[7] http://code.google.com/p/information-artifact-ontology/ (last accessed April 13th 2012)
[8] Rosse C, Mejino JLV. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. Journal of Biomedical Informatics. 2003;36(6):478-500.
[9] http://www.ihtsdo.org/snomed-ct/ (last accessed April 13th 2012)
[10] Brinkman R, Courtot M, Derom D, Fostel J, He Y, Lord P, Malone J, Parkinson H, Peters B, Rocca-Serra P, Ruttenberg A, Sansone S, Soldatova L, Stoeckert C Jr, Turner J, Zheng J; OBI consortium. Modeling biomedical experimental processes with OBI. J Biomed Semantics. 2010;1 Suppl 1(22):S7.
[11] Gostev M, Fernandez-Banet J, Rung J, Dietrich J, Prokopenko I, Ripatti S, McCarthy MI, Brazma A, Krestyaninova M. SAIL--a software system for sample and phenotype availability across biobanks and cohorts. Bioinformatics. 2011;27(4):589-91.
[12] http://www.euengage.org/ (last accessed April 13th 2012)
[13] http://www.elixir-europe.org/ (last accessed April 13th 2012)
[14] http://www.datashaper.org/ (last accessed April 13th 2012)
[15]. Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, Zhukova A, Brazma A, Parkinson H. Modeling sample variables with an Experimental Factor Ontology. Bioinformatics 2010; 26(8):1112-1118.