21st International Congress on Modelling and Simulation, Gold Coast, Australia, 29 Nov to 4 Dec 2015
www.mssanz.org.au/modsim2015

PROV and Real Things

Simon J.D. Cox [a] and Nicholas J. Car [a]

[a] *Land & Water Flagship: CSIRO, Melbourne Vic and Brisbane, Qld, Australia*
Email: simon.cox@csiro.au

Abstract: The PROV data model is becoming accepted as a flexible and robust tool for formalizing information relating to the production of documents and datasets. Provenance stores based on the PROV-O implementation are appearing in support of scientific data workflows. However, the scope of PROV does not have to be limited to digital or information assets. For example, specimens typically undergo complex preparation sequences prior to actual observations and measurements, and it is important to record this to ensure reproducibility and to enable assessment of the reliability of data produced. PROV provides a flexible solution, allowing a comprehensive trace of predecessor entities and transformations at any level of detail. In this paper we demonstrate the use of PROV for describing specimens managed for scientific observations. Two examples are considered: a geological sample which undergoes a typical preparation process for measurements of the concentration of a particular chemical substance, and the collection, taxonomic classification and eventual publication of an insect specimen. We briefly compare PROV with related work.

Keywords: *Provenance, PROV, sampling, specimen*

1. INTRODUCTION

The concept of provenance was developed in the art, museums and archives community, referring to a record of the history of an artefact. Tracing the history of ownership and custodianship is an important means of determining the authenticity of rare or unique items.

More recently, the same term has been applied to the lineage of information objects - particularly datasets, imagery, etc. These may have a complex history with multiple data-processing steps, the details of which are important in evaluating the quality or fitness for a particular purpose, and also for establishing reproducibility which is a core tenet of empirical science. In this setting, a 'provenance trace' can be seen as the record of an instantiated 'workflow'. The W3C PROV model [6,9] harmonizes a number of earlier treatments (in particular PML and OPM) and is becoming accepted as the basis for formalizing information relating to the production of documents and datasets. Provenance stores based on the PROV-O implementation, such as PROMS [1], are appearing in support of scientific data workflows.

In the context of technical and scientific collections, where specimens (biological, water, soil and rock) are managed to support subsequent observations, chain of custody concerns do arise, particularly in relation to forensic applications, and also where there are financial implications from results of observations on samples, such as assays on mineral exploration specimens. But the key feature of technical and scientific specimens is the preparation process, generating a sequence of samples - some of which only exist temporarily - which are related through various processing activities, in support of the observational requirements. There is enormous variety in the process-chains related to these, both between and within disciplines. In fact, the design of new sequences is a key activity in empirical science. The PROV ontology - which abstracts all possible processing chains into a single high level model whereby the production and transformation of *Entities* is through time-bounded *Activities*, under the influence or control of *Agents* - appears to provide a framework that can satisfy all of the relevant concerns, either directly or through minor specializations within a basic framework.

While there is little precedent in using PROV for non-information assets, this is not specifically excluded by either the specification or any aspects of the logic in the ontology, and is obliquely referred to in at least one example in the specification. The appeal of using PROV for specimens is (i) it has a flexible and extensible model that can capture any sample collection and preparation scenario; (ii) descriptions of samples and specimens can be stored and accessed through systems integrated with other aspects of the scientific workflow; (iii) being an RDF-based technology it can be accessed and processed by a growing set of standard tools, including SPARQL for low-level access; (iv) being an OWL-based ontology, it also supports rigorous reasoning; (v) general compatibility with Linked Data principles.

The description of specimens has previously been described in the Biological Collections Ontology (BCO) [13]. In common with many vocabularies from the biomedical community, BCO is aligned to the Basic Formal Ontology (BFO) [7]. For chain-of-custody semantics, OntoPedigree [11,12] has been developed to support supply-chain applications. The OGC Observations and Measurements (O&M) [2,3,8] established a basic model for sampling features, including a specialization for specimens. In an OWL implementation of that model, Cox [4] adopted PROV to provide flexible support of specimen-preparation descriptions, and also as a convenient 'upper ontology' for alignment and analysis of the O&M model, and to assist in mapping with other observation ontologies.

In this paper we step back and limit ourselves to only PROV classes and properties. We demonstrate that PROV may be used to describe the history and preparation of physical specimens through two worked examples. The first describes a sample of rock which undergoes a multi-step preparation process prior to some specific measurements. The second describes the generation of a report classifying an organism collected in the field, involving various intermediate material and information entities.

2. OVERVIEW OF PROV

PROV defines a model for building representations of the entities, people and processes involved in producing a piece of data or thing in the world [6].

In the core PROV model there are just three top-level classes: Activity, Entity, and Agent. Activities result in changes to Entities, including their creation. An Agent is involved in, or has responsibility for, the activity. PROV provides a terminology for the relationships between individuals from these classes, with nine generic properties as shown in Figure 1. Entities are typically information artefacts, such as datasets, representations of these in prose, graphics and tables, and reports and papers. Agents are typically people, organizations, and

software. Activities typically concern data processing and the preparation of reporting artefacts. Activities are usually associated with an Agent, and may use existing Entities as input. Activities are ordered by the timestamps associated with their beginning and end, but also logically by the requirement for outputs of one activity to be available as inputs to the next. A chain of Entities may be defined directly through the wasDerivedFrom property, but also may be inferred by the logical or temporal ordering of Activities and their inputs and outputs.

PROV itself provides additional classes and properties, many of which specialize the ones described here, and further extensions may be made to support specific communities and scenarios, but Figure 1 shows the essential model.

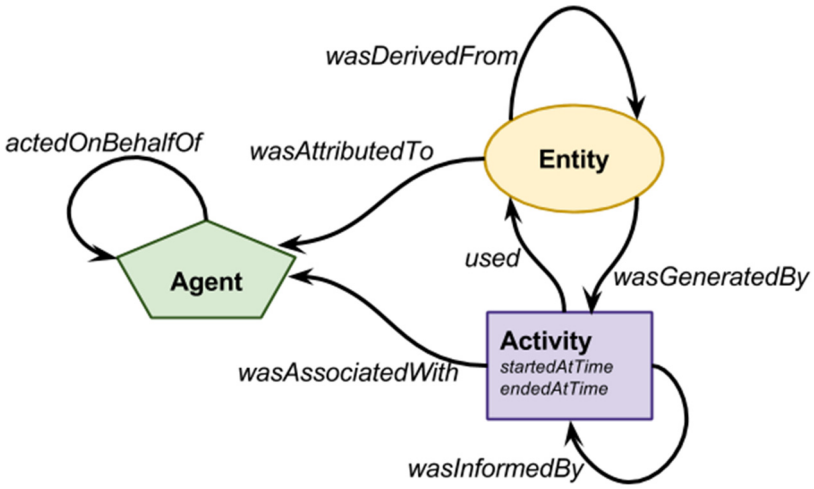

Figure 1. High-level view of the PROV model, showing the three core classes and the generic properties that relate them.

3. WORKED EXAMPLES

3.1. Geological Sample

We have taken the story of the retrieval of a geological sample, its preparation for observation, and the measurements made on it[1], and described this using only the core PROV model (see Figure 2). The basic idea is that the thing in the field (a geologic unit, bed, formation, etc) and each physical sample is classified as a prov:Entity. Each processing step, from initial retrieval, through splitting, crushing and sieving, is a prov:Activity that results in one or more new specimens. Each step (prov:Activity) uses the previous specimen (prov:Entity) as input, and some machine or process, operated by a scientist or technician, all of which are classified as prov:Agents in this context. In the final step, two analyses (prov:Activity) are carried out by further agents, each following a specified protocol (prov:Plan - a specialized prov:Entity), resulting in two datasets (prov:Entity), which report estimates of the carbonate proportion within fractions defined by grain-size.

Note that the provenance record or trace is retrospective: the relationships from each prov:Entity are backward to the prov:Activity that generated it, and optionally to the prior prov:Entity(s) from which it was derived. In turn, each prov:Activity has properties that point backwards to the inputs, protocols, and machines and their operators involved. Thus, a provenance trace is the realization of a planned workflow, with each element realized as a concrete instance of the relevant type.

In practice, entities intermediate between the specimen retrieved from the field and the reported measurements are usually of little interest outside the laboratory where the work was carried out, so identifiers for all specimens except the initial one are not generally published. Nevertheless, it is useful to explicitly acknowledge the existence of the intermediate specimens, as shown in the provenance trace in Figure 2, as they may be required for quality assessment and reproducibility purposes. This has been a fraught area in routine geochemistry [5]. In the commercial assay labs there is a rich, but non-standardized, practice and terminology of ‘splits’, ‘repeats’, ‘duplicates’, ‘replicates’, etc., often denoted primarily through a complex ‘sample naming convention’, which is unfortunately local to each lab. Hence, a systematic approach, which is also flexible to

[1] Thanks to Keith Sircombe of Geoscience Australia for providing the story upon which this example is based.

accommodate varying and novel processing chains, is a significant opportunity for rationalization and clarification.

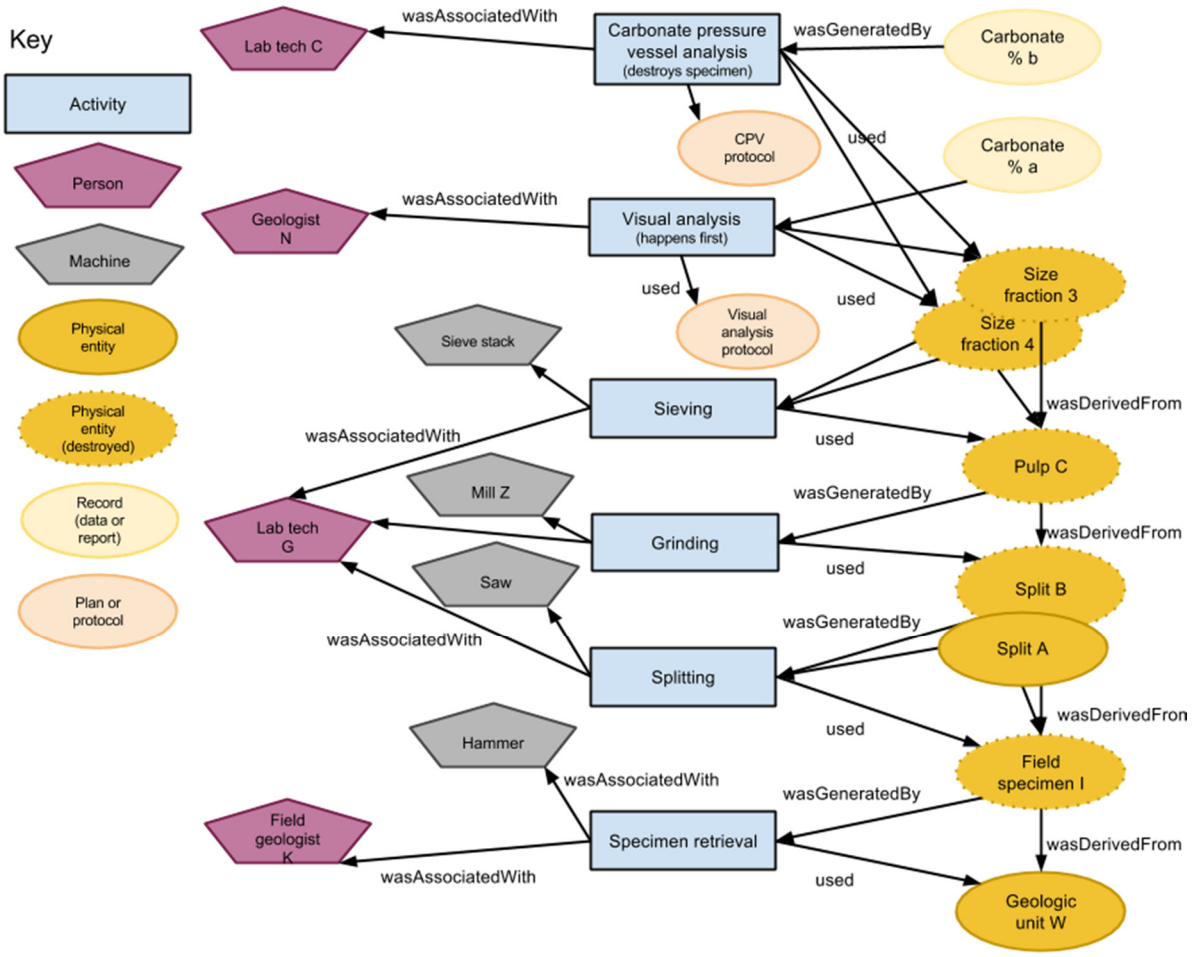

Figure 2. Agents (people and machines), activities, and entities in a provenance trace describing the retrieval, preparation and analysis of a geological specimen for carbonate content. The final outputs of the process are datasets (top right) which record the results of observations whose intention is to characterize the real-world Geologic Unit (bottom right), but there is a sequence of intermediate specimens involved. Some intermediate specimens only exist temporarily as they are consumed (destroyed) by the subsequent activity.
The RDF source is available from https://github.com/CSIRO-LW-LD/prov/tree/master/real-things

3.2. Biological sample

The second example considers the collection, taxonomic classification and eventual publication of an insect specimen (see Figure 3). This trace does not show the same level of detail relating to sample preparation, but includes more of the story around data and publication, and the use of prior data.[2] While the specific data used to make species classification within different genera/families/orders/classes of biological samples, the general process presented here is applicable to a wide range of species classifications.

Quite a range of prov:Entity elements are present in this example from a subsample of the physical specimen to the amorphous "taxonomic data (papers etc.)" used to represent information needed by the taxonomist to classify the specimen according to species.

As per the geological sample shown in Figure 2, the physical sample – specimen - is classified as a prov:Entity. All actions taken to prepare the specimen for preservation, to sample it or to generate data from it are considered

[2] Thanks to Catherine Car of the Western Australian Museum for providing a story that is the basis of this example.

prov:Activities. The "taxonomic protocols" governing species identification of the specimen based on morphological and other measurements is represented as a prov:Plan as it guides the classification Activity.

Unlike the geological sample example where intermediate samples of the specimen are not necessarily of interest and often not preserved, tissue samples of insect specimens can be of great interest and stored as a subsample of the specimen in perpetuity, especially of the specimen becomes a holotype for a newly identified species.

One area that may require further thought is the relationship between sample/specimen and its representation as a single or separate prov:Entity. The "Preservation and preparation" Activity involves killing the live specimen and preserving it dry or in alcohol. In this example, the live and dead specimen is identified as two different Entities.

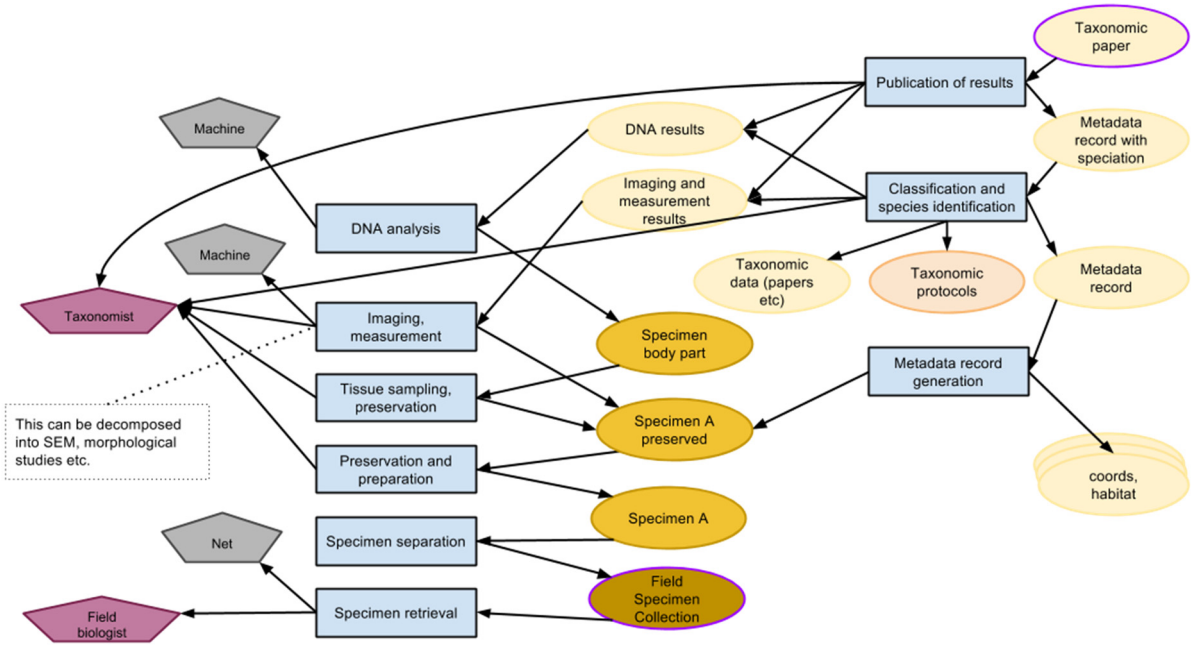

Figure 3. Agents, activities, and entities in a provenance trace describing the collection, and preparation of an insect specimen and the determination of its taxon. The final output of the process is a paper (top right) which announces the results of a taxonomic classification based on combining morphological and genetic observations. The trace includes both material entities (dark yellow) and intermediate datasets (pale yellow) as well as ancillary data inputs (prior taxonomic data), intermediate datasets (imaging, DNA results) and outputs (formal metadata records for a data repository). The RDF source is available from https://github.com/CSIRO-LW-LD/prov/tree/master/real-things

4. DISCUSSION

A number of questions must be resolved when applying PROV to real things. For example, prov:Agent has the built-in subclasses prov:Person, prov:Organization, and prov:SoftwareAgent which indicates the expected scope of prov:Agent. Inanimate objects like machines and scientific apparatus play a similar role in relation to the generation and transformation of real things as software does in relation to data and information resources, so we have also classified these as prov:Agents. They do not have 'agency' in the sense of 'free-will' (neither do software agents) but they do effect the change of the entities involved. Alternatively, each of these could be thought of as merely a prov:Entity that is 'used' in the prov:Activity, and from the point of view of asset management software and hardware are clearly prov:Entities. Note that classifying them as prov:Agents is not inconsistent with them also being prov:Entities, as there is no axiom or set of axioms in PROV that creates a conflict.

On the other hand, prov:Activity and prov:Entity are disjoint classes - there is a fundamental distinction made between events that transform things, and entities that persist. This is a key aspect of the view presented here, which focuses on the transformation steps, so the prov:Activities carry most of the properties that tell the story. This view is also consistent with BFO [7], in which the distinction between 'continuants' and 'occurrents' is

the most fundamental division. Hence, it is relatively straightforward to compare the model presented here with BCO [13], which uses BFO:Process and BFO:MaterialEntity as the root classes for sample preparation processes and samples, respectively. However, the agents responsible for the processes, and therefore get credit for creating the entities, are not a core class in BCO (or BFO) so the alignment is incomplete.

In the work reported here we limited ourselves to the core PROV classes and properties, in order to demonstrate that the PROV pattern supports the application. Note that the PROV model also includes specializations of the core classes (Agents, Entities and Activities) as well as a hierarchy of linking classes headed by "prov:Influence" to group information about an activity, like the specific roles of agents involved. These would be used in a more complete implementation. However, the specializations built in to the PROV standard are optimised primarily for document and report preparation, so additional specializations will be needed for real-things applications, usually through sub-classes and sub-properties of PROV. For example, the IGSN registration model enumerates a set relationships (IsCitedBy, IsPartOf, HasPart, IsReferencedBy, References, IsDocumentedBy, Documents, IsCompiledBy, Compiles, IsVariantFormOf, IsOriginalFormOf) not all of which can be mapped to the PROV properties that link entities. In Table 1 we show a number of potential specialization for generic and geological sample preparation. Other domain-based specializations are also likely.

Table 1. Notional specializations of PROV to support sample preparation descriptions. Sub-classes and Sub-properties indicated with "←", Geological applications in brown

prov:Entity ← :PhysicalEntity ← :Specimen prov:Entity ← prov:Plan ← :SamplingProtocol
prov:Agent ← :SampleProcessingSystem ← :GrindingSystem, :PolishingSystem, :DissolvingSystem, :FusingSystem prov:Agent ← :SampleRetrievalSystem ← :FieldSamplingSystem prov:Agent ← :SubSamplingSystem ← :BiasedSplittingSystem ← :SizeSeparationSystem , :DensitySeparationSystem, :MagneticSeparationSystem prov:Agent ← Instrument , Sensor
prov:wasAssociatedWith ← :wasControlledBy, :wasSponsoredBy, :wasRequestedBy
prov:wasDerivedFrom ← :unbiasedSplitFrom, :biasedSplitFrom
prov:hadPrimarySource ← :fieldSpecimen

In Table 2 we sketch an initial mapping of PROV, and some specializations for sample preparation, with BCO and BFO. Since BCO had already focused on real things, the alignment is straightforward.

Table 2. Alignment of PROV and the sampling extension with BCO.

BFO:Process ← prov:Activity
BFO:IndependentContinuant ↔ prov:Entity ← BFO:MaterialEntity ↔ :PhysicalEntity ← BCO:MaterialSample ↔ :Specimen

The issue of sample identity is important to help correlate results of data generated about the same field specimen from different laboratories, and reported in different publications. The International Geosample Number (IGSN) system [10] has been developed to address this within the geosciences, through which specimens are registered and assigned a unique identifier, which is then cited in all formal publications and reports. The PROV-O implementation of PROV [9] uses URIs to identify all resources in a provenance trace. IGSNs are also accessible as URIs, so if registered in the IGSN system, that identifier would be used for "Field specimen I" in the example shown in Figure 2 (and attached to the specimen using a QR-code or bar-code). There are similar efforts towards sample identifiers in other disciplines. The formalized approach to provenance records described here is therefore highly compatible with existing community standards.

Chain-of-custody applications have not been considered in this paper. It might be possible to model this using PROV through a sequence of custody-transitions modeled as Activities. However, this would imply a new identified Entity for each custody-state. Alternatively, a 'Custody' activity, which does not involve any change

in the Entity, might be defined, but other aspects of the application of PROV to this case are not at all clear. It appears that the PROV model is not optimised for this kind of provenance application, and an alternative vocabulary must be used or designed.

5. CONCLUSION

We have demonstrated the application of the W3C PROV ontology for description of physical specimens used for scientific observations. The core PROV model is well suited to this application, with no changes to the core model required for quite sophisticated descriptions of two divergent cases, running right through from field data collection to publication. Nevertheless, domain-specialization might support more sophisticated applications. Managing the description of real things using the PROV formalization provides the opportunity to join up the internet of things and the semantic web.

REFERENCES

[1] Car, Nicholas J.; Stenson, Matthew Paul; and Hartcher, Michael, "A Provenance Methodology And Architecture For Scientific Projects Containing Automated And Manual Processes" (2014). *International Conference on Hydroinformatics.* Paper 57. http://academicworks.cuny.edu/cc_conf_hic/57 (accessed July 31, 2015).

[2] S.J.D. Cox, Observations and Measurements – Part 2 - Sampling Features - OGC 07-002r3, Open Geospatial Consortium, Wayland, Mass., 2007. http://portal.opengeospatial.org/files/22467.

[3] S.J.D. Cox, Geographic Information - Observations and Measurements (OGC Abstract Specification Topic 20) (same as ISO 19156:2011), OGC 10-004r3. (2011) 54. http://portal.opengeospatial.org/files/41579 (accessed September 16, 2014).

[4] S.J.D. Cox, Ontology for observations and sampling features, with alignments to existing models, Semant. Web J. (2015) Submitted. http://www.semantic-web-journal.net/content/ontology-observations-and-sampling-features-alignments-existing-models (accessed July 24, 2015).

[5] S.J.D. Cox, A. Dent, S. Girvan, R. Atkinson, I. Whitehouse, C. Legg, Using the Assay Data Exchange standard with WFS to build a complete minerals exploration data-transfer chain, in: Proc. Int. Assoc. Math. Geol. Gen. Assem., 2006.

[6] Y. Gil, S. Miles, PROV Model Primer, (2013). http://www.w3.org/TR/2013/NOTE-prov-primer-20130430/#intuitive-overview-of-prov (accessed July 27, 2015).

[7] P. Grenon, B. Smith, SNAP and SPAN: Towards Dynamic Spatial Ontology, Spat. Cogn. Comput. 4 (2004) 69–103. doi:10.1207/s15427633scc0401_5.

[8] ISO/TC-211, ISO 19156:2011 - Geographic information -- Observations and measurements, (2011). http://www.iso.org/iso/catalogue_detail.htm?csnumber=32574 (accessed February 4, 2014).

[9] T. Lebo, S. Sahoo, D. McGuinness, PROV-O: The PROV Ontology, (2013). http://www.w3.org/TR/prov-o/ (accessed February 13, 2014).

[10] K.A. Lehnert, J. Klump, R.A. Arko, S. Bristol, B. Buczkowski, C. Chan, et al., IGSN e.V.: Registration and Identification Services for Physical Samples in the Digital Universe, in: *AGU Fall Meet.*, American Geophysical Union, 2011: p. IN13B–1324. http://abstractsearch.agu.org/meetings/2011/FM/IN13B-1324.html (accessed July 30, 2015).

[11] M. Solanki, C. Brewster, Consuming Linked data in Supply Chains: Enabling data visibility via Linked Pedigrees, in: Fourth Int. Work. Consum. Linked Data, 2013: p. 13. http://ceur-ws.org/Vol-1034/SolankiAndBrewster_COLD2013.pdf (accessed July 24, 2015).

[12] M. Solanki, C. Brewster, OntoPedigree: A content ontology design pattern for traceability knowledge representation in supply chains, Semant. Web J. (2015) in press. http://www.semantic-web-journal.net/content/ontopedigree-content-ontology-design-pattern-traceability-knowledge-representation-supply-2 (accessed July 24, 2015).

[13] R.L. Walls, J. Deck, R. Guralnick, S. Baskauf, R. Beaman, S. Blum, et al., Semantics in support of biodiversity knowledge discovery: an introduction to the biological collections ontology and related ontologies., PLoS One. 9 (2014) e89606. doi:10.1371/journal.pone.0089606.