*Application of Semantic Technology in Biodiversity Science*
*Anne E. Thessen (Ed.)*
*ISBN 978-3-89838-733-0*

# Chapter 2

# Darwin Core as a Vocabulary for Expressing Biodiversity Data as RDF

Steven J. BASKAUF[a] and Joel SACHS[b]
[a]Vanderbilt University Department of Biological Sciences, Nashville, Tennessee, USA
[b]Agriculture and Agri-Food Canada, Ottawa, Ontario, Canada

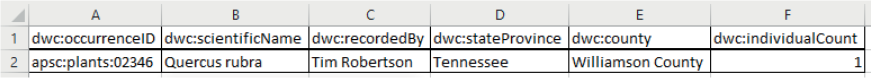
Darwin Core is a glossary of terms intended to facilitate the sharing of biodiversity occurrence data and related information, such as specimen metadata, checklists, determination histories, and taxonomy. It is widely used by museums and herbaria to share specimen data with aggregators through fielded text files. This chapter discusses efforts to use Darwin Core terms to express these data as RDF. It describes the history and motivation of the Darwin Core RDF Guide, presents some of the difficulties involved in using Darwin Core in RDF, and gives examples that demonstrate the challenges that still remain for the community to address.

## 1. Background

In 2005, Biodiversity Information Standards (TDWG) embarked on a mission to establish an umbrella architecture for TDWG standards. The goal of this effort was to enable data integration among providers of heterogeneous datasets in cases where the ultimate use and users were unknown[1]. By 2007, architecture development was focused on creation of a high level OWL ontology, development of an XML-based exchange protocol (TAPIR[2]), and on adoption of Life Science Identifiers (LSIDs), a globally unique identifier system[3]. Within several years, it became apparent that this approach was not working. Developing a TDWG ontology was too complex and time-consuming, consensus was elusive, data transfer by the TAPIR protocol was too slow, and implementation of LSIDs carried too heavy of a technical burden for it to be widely adopted.

By 2009, a simpler system had evolved to allow effective data transmission. The Darwin Core (DwC) Standard[4] provided a vocabulary of terms that could be used to describe occurrences and taxa, the core resources of interest to TDWG. Darwin Core Archives, a system for transmitting records via compressed delineated text files, was highly efficient[5]. A consensus resolution to the problem of globally unique identifiers was not attained, and providers used ad hoc solutions such as "Darwin Core Triplets" or UUIDs. Although this system facilitated automated harvesting of provider data, it did not satisfy the original goal of enabling integration of heterogeneous data.

From 2009-2011, increased interest in Linked Open Data[6] within TDWG resulted in the suggestion that a system for machine-mediated data integration might be built around Darwin Core using standard Linked Data technologies. In 2011, TDWG chartered an RDF/OWL Task Group to explore best practices for the use of RDF in the biodiversity informatics realm. It became quickly clear that the semantics of Darwin Core, while sufficient for representing data in spreadsheets, were too ambiguous for the semantic web, where a human user cannot be counted on to interpret intended meaning from surrounding context. Some of the issues related to distinguishing between resources and descriptions of resources that are common to all domains migrating to Linked Data. Others were particular to biodiversity informatics, and its historical grounding in natural history collections. Workshops were held in 2013 and 2014 to address these issues, and sufficient progress was made to enable the Task Group to deliver a Darwin Core RDF guide, which was adopted as an addition to the Darwin Core standard[7].



|  | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | dwc:occurrenceID | dwc:scientificName | dwc:recordedBy | dwc:stateProvince | dwc:county | dwc:individualCount |
| 2 | apsc:plants:02346 | Quercus rubra | Tim Robertson | Tennessee | Williamson County | 1 |

**Figure 1.** An occurrence record expressed as "Simple Darwin Core".

## 2. How the RDF guide makes Darwin Core usable in Linked Data

The simplest way to represent metadata as Darwin Core is in the form of a single spreadsheet or CSV table. Figure 1 shows an imaginary occurrence record for red oak.

Each column header of the table contains a Darwin Core property from the namespace http://rs.tdwg.org/dwc/terms/ (commonly abbreviated dwc:), and the cells of the table contain the values of those properties. A particular row of the table contains information about a particular instance of an occurrence. Since the column headers contain abbreviated IRIs (CURIEs[8]), it would be a straightforward matter to convert the data in this table directly into an RDF graph. Graphically, the metadata for the row would look like Figure 2.

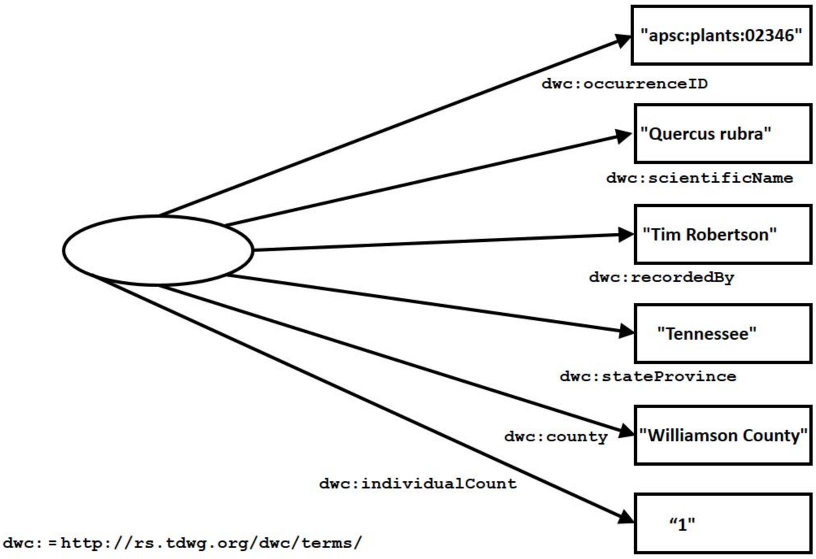


**Figure 2.** The occurrence record of Figure 1 converted directly into RDF.

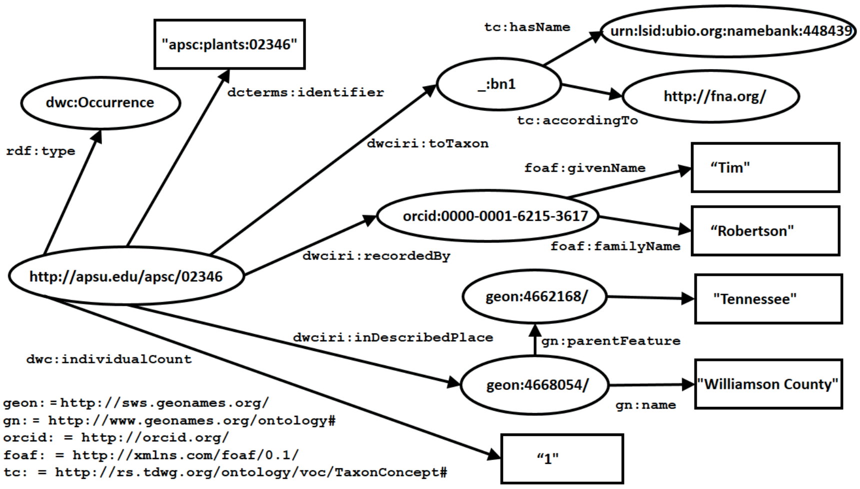


**Figure 3.** The occurrence record of Figure 1 converted into RDF in the spirit of Linked Data.

From this diagram, we can easily see several features of the resulting RDF graph. First, the occurrence instance itself is not named with a IRI - it is a blank node. Second, the graph is very "flat", meaning that the occurrence node has no more than a single edge connecting it to any other entity. Third, the object resource of every triple is a literal[1] (indicated by the rectangle). In some cases, the object resource is best represented by a literal, such as the identifier "apsc:plants:02346" or the number "1". But many of the resources would be better represented by non-literal objects, since they represent real-world entities (place, person) or abstract concepts (a taxon) to which URIs have (or could be) assigned.

The graph in Figure 3 differs from Figure 2 in several significant ways. The focal occurrence instance is identified by a IRI. For triples where the subject is the occurrence instance and the object is a non-literal resource (ovals), the object is denoted by an IRI or by a

[1] In RDF, the object of a triple can be a blank node, an IRI, or a literal. Examples of literals are "19", "July 22, 2011", "Tim Robertson", "London England". If the goal is data integration, best practice is to use IRIs where possible, and to use literals only for representing things like numbers and dates. Because RDF does not permit literals as subjects of triples, literals are always sinks in an RDF graph (i.e., they are nodes with no outgoing edges).

blank node identifier. In many cases, these non-literal resources are themselves the subjects of additional triples that describe their properties. As a result, the graph is not “flat”, but rather is a network. In the spirit of Linked Data, many of the linked non-literal objects are denoted by IRIs minted by authorities other than the source that created the occurrence record.

The primary purpose of the Darwin Core RDF Guide was to facilitate the transition of flat, literal-object based RDF that might be directly exported from a relational database or spreadsheet into RDF that describes a more complex graph containing links to resources described more fully by other providers. That goal was achieved by establishing conventions for handling several idiosyncrasies that result from the fact that Darwin Core was designed with flat tables in mind.

Listed under each class in Darwin Core, there is a term whose local name ends in “ID” and whose purpose is to indicate an identifier for an instance of that class (e.g., dwc:occurrenceID in Figures 1 and 2[9]). The ID terms can be used as either foreign keys or primary keys in Darwin Core records. For example, in an occurrence record, the occurrenceID field would be used to provide an identity for the record itself, while the identificationID field would be used to specify an identification record related to the occurrence. In RDF, however, triples should not depend on context for their semantics, so the RDF Guide specifies that the well-known term dcterms:identifier should be used to indicate the identifier associated with the subject resource and the standard term rdf:type should be used to link to the IRI of the class of which the subject is an instance (Fig. 3). This is standard practice in the Linked Data community.

Term definitions in generic Darwin Core are not specific about what kind of literal should be used to denote objects like people, places, and taxa. The value might be a name, an identifier (including an IRI), or a string representing a controlled vocabulary term. For each instance, an aggregating client would probably be required to do some sort of string cleaning and matching with a list of known values in order to determine whether the object resource is already a known entity. There is also the possibility that the value is not unique to a particular resource. This places a large burden on the consumer. For example, in Figure 2, the value of dwc:recordedBy could denote any of several persons named “Tim Robertson” and would not match with other records that might refer to the same person but use different forms of his name. The RDF Guide allows for terms in the generic dwc: namespace whose values denote non-literal resources to be used with literals in this way. However, the Guide creates alternative analogs of those terms whose values are expected to be IRIs. Those analogs have the same local name as the generic terms, but are placed in the namespace http://rs.tdwg.org/dwc/iri/ (commonly abbreviated as dwciri:). The burden on the aggregating client discovering a triple containing a dwciri: predicate should be less that a triple containing a predicate from the dwc: namespace, since the value of a dwciri: predicate should be a globally unique IRI. If that IRI is dereferenceable, the client may be able to acquire additional information about the object resource that will make its identity clear, and which may also lead the client to discover useful related information. In Figure 3, the value of dwciri:recordedBy is an ORCID ID that is unique and unambiguous, and that will provide additional information when dereferenced. If a data provider or aggregator makes the effort to replace an ambiguous dwc: term literal value with an unambiguous dwciri: value, that effort need only be made once, rather than requiring the disambiguation to be done by every user.

Darwin Core defines a number of sets of terms that describe a hierarchy of literal values that, as a set, provide an unambiguous reference to some resource. For example, terms such as dwc:county, dwc:stateProvince, dwc:country, and dwc:continent can be used to describe the geographic hierarchy into which a location falls (Fig. 2). Such a hierarchy of terms is convenient for searching (hence the name “convenience terms”), but using those terms requires defining the hierarchy in every record of the dataset. The RDF Guide provides several new terms that facilitate linking to the IRI of the lowest level of a standardized hierarchy. For example, in Figure 3, the term dwciri:inDescribedPlace links to the IRI for Williamson County, Tennessee. Using this IRI eliminates ambiguity about which Williamson County is denoted and by dereferencing the IRI, a user can discover that Tennessee is part of the United States and that the United States is part of North America without having to repeat that information in every record.

## 3. Barriers to implementation

### 3.1. Lack of IRIs

In addition to the difference in complexity between the graphs represented in Figures 1 and 2, another obvious difference is that the focal resource in Figure 1 is a blank node, whereas in Figure 2 that resource is identified with an HTTP IRI. A key tenet of the Linked Data paradigm is that resources should be identified by HTTP IRIs. But who should be responsible for minting those IRIs? One approach is to leave the minting to some centralized authority. The ARCTOS database[10] mints globally unique IRIs for all specimens from collections whose data it aggregates based on the specimen's Darwin Core Triple. For example, the IRI `http://arctos.database.museum/guid/MVZ:Mamm:165861` dereferences to a record from the Museum of Vertebrate Zoology. Another approach is to delegate the responsibility of minting IRIs to the institution holding the described resource. This approach was taken in the Stable Identifiers initiative of the Consortium of European Taxonomic Facilities (CETAF)[11]. The initiative describes best practices for identifier creation, but leaves it up to participating institutions to implement those practices.

If an institution does not mint its own HTTP IRI identifiers or participate in an aggregation effort that mints IRIs for it, then there will be no well-known IRI for the resources described in that institution's records. Of course, anyone could "make up" an IRI (as we did in Fig. 3), but such an IRI would not dereference, nor would it correspond to another ad hoc IRI for the same resource minted by someone else.

This lack of HTTP IRI identifiers for core resources in the biodiversity informatics realm is a serious impediment to implementation of Linked Data principles. Given that there are many possible technical solutions for the problem of generating stable HTTP IRIs, this barrier is social, not technological.

### 3.2. Inconsistent graph models

RDF is fundamentally a system for expressing data as a graph. So although several providers may agree to follow the Darwin Core RDF Guide, there is no guarantee that the graphs they create could be

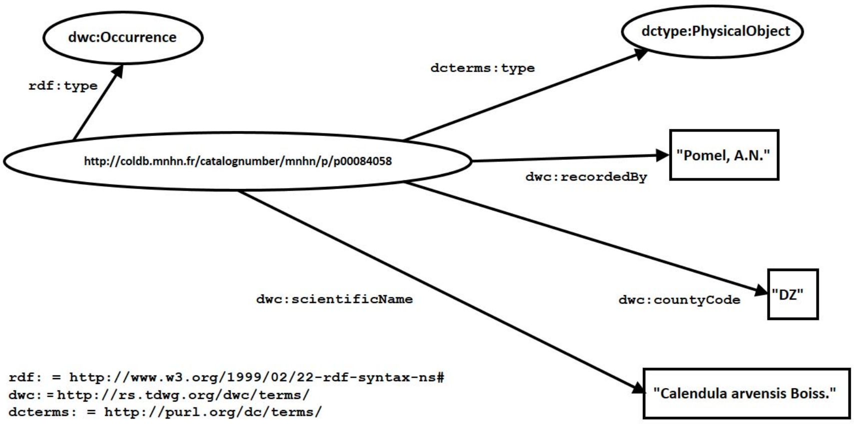

dwc:Occurrence
dctype:PhysicalObject
dcterms:type
rdf:type
http://coldb.mnhn.fr/catalognumber/mnhn/p/p00084058
"Pomel, A.N."
dwc:recordedBy
dwc:scientificName
dwc:countyCode
"DZ"
rdf: = http://www.w3.org/1999/02/22-rdf-syntax-ns#
dwc: =http://rs.tdwg.org/dwc/terms/
dcterms: = http://purl.org/dc/terms/
"Calendula arvensis Boiss."

**Figure 4.** Graphical representation of some RDF triples related to the occurrence record `http://coldb.mnhn.fr/catalognumber/mnhn/p/p00084058` from the Muséum national d'Histoire naturelle, Paris.

productively merged if they based their RDF on different graph models.

Figures 4 and 5 illustrate RDF graphs of two specimen-related occurrences from different providers.

Although both graphs provide essentially the same metadata about the occurrences, the graphs are radically different in their degree of complexity. In Figure 4, the properties that describe the collector name, the country in which the collection was made, and the scientific name are all linked directly to the occurrence record. Although it is not explicitly declared to be an instance of dwc:PreservedSpecimen, it is clear by the dcterms:type declaration of dctype:PhysicalObject that the data providers consider the occurrence to be the same entity as the specimen. (It should be noted that this usage is not consistent with the definition of dwc:Occurrence - An existence of an Organism (sensu http://rs.tdwg.org/dwc/terms/Organism) at a particular place at a particular time.) In contrast, in Figure 5, none of these metadata are linked directly to the occurrence. Rather, the collector name is linked to a foaf:Person instance, the country code is linked to a dcterms:Location instance, and the scientific name is linked to a dcterms:Location instance. Additionally, in Figure 5 the actual

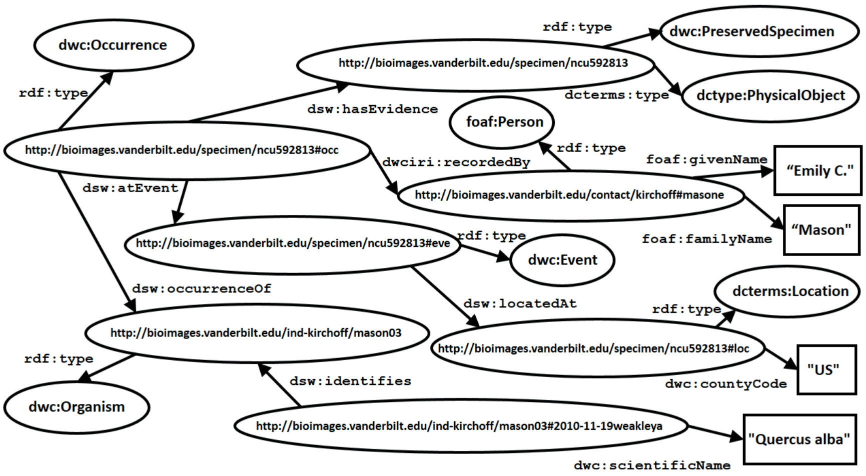

dwc:Occurrence
rdf:type
dwc:PreservedSpecimen
http://bioimages.vanderbilt.edu/specimen/ncu592813
dcterms:type
dctype:PhysicalObject
rdf:type
dsw:hasEvidence
foaf:Person
http://bioimages.vanderbilt.edu/specimen/ncu592813#occ
rdf:type
foaf:givenName
"Emily C."
dwciri:recordedBy
http://bioimages.vanderbilt.edu/contact/kirchoff#masone
dsw:atEvent
"Mason"
http://bioimages.vanderbilt.edu/specimen/ncu592813#eve
rdf:type
foaf:familyName
dwc:Event
dsw:occurrenceOf
dsw:locatedAt
dcterms:Location
rdf:type
http://bioimages.vanderbilt.edu/ind-kirchoff/mason03
http://bioimages.vanderbilt.edu/specimen/ncu592813#loc
rdf:type
"US"
dsw:identifies
dwc:countyCode
dwc:Organism
http://bioimages.vanderbilt.edu/ind-kirchoff/mason03#2010-11-19weakleya
"Quercus alba"
dwc:scientificName

**Figure 5.** Graphical representation of some RDF triples related to the occurrence record `http://bioimages.vanderbilt.edu/specimen/ncu592813#occ` from University of North Carolina Herbarium.

preserved specimen is considered a distinct resource that provides evidence for the occurrence instance, but which is not considered equivalent to it.

There is nothing that would prevent merging these two graphs, but constructing a simple query that would work for both datasets would be difficult. For example, using the graph model of Figure 4, finding occurrences from Algeria could be accomplished with this query:

```
SELECT DISTINCT ?occurrence
WHERE {
        ?occurrence dwc:countryCode "DZ".
        }
```

Using the graph model of Figure 5, finding occurrences from the United States could be done with this query:

```
SELECT DISTINCT ?occurrence
WHERE {
        ?occurrence dsw:atEvent ?event.
        ?event dsw:locatedAt ?location.
?location dwc:countryCode "DZ".
}
```

At first glance, the graph model of Figure 5 seems unnecessarily complex. However, the dataset of which it is a part is designed to allow for multiple determinations per organism, multiple occurrences per organism, multiple forms of evidence serving as evidence for a single occurrence, multiple events at one location, and multiple occurrences at one event. None of these many-to-one relationships can be easily described using the model of Figure 4.

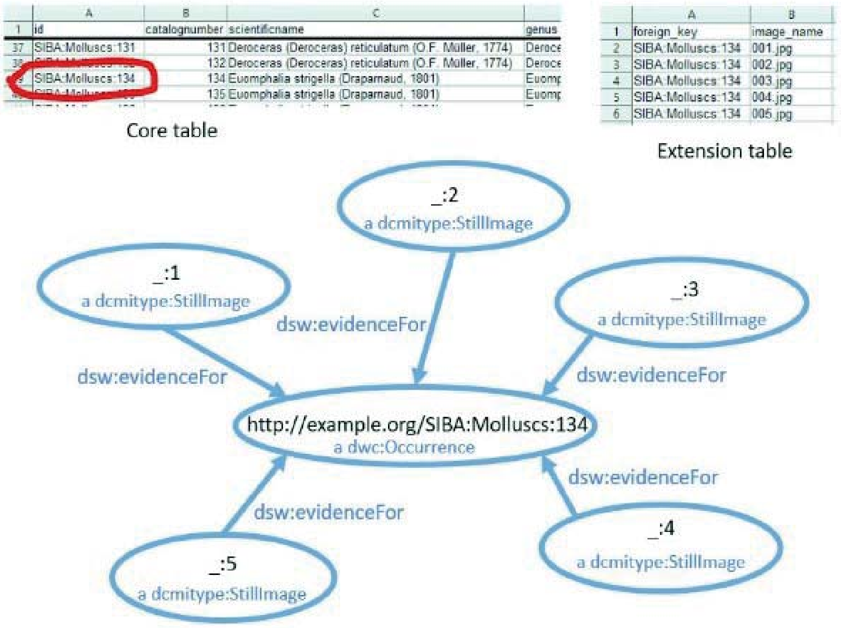
The main point here is that the graph model that is right in a particular circumstance is one that is just complex enough to satisfy the use cases that are important to users of that model. Thus, it is not really possible to say what the "right" graph model is without first determining who the potential users are, and what use cases are important to those users.

### 3.3. Variety of datasets represented in the wild

We can get some sense of the scope of use cases that interest members of the biodiversity informatics community by examining the structure of a variety of publically available datasets. Clearly, there are many preserved specimen datasets represented in a single table that can be modeled simply like the example in Figure 4. However, the structure of a Darwin Core Archive allows for a core table to be linked to multiple extension tables in an organizational pattern that has been called a "star schema"[12]. If represented as an RDF graph, such a pattern would be composed of a node of a central resource class linked by a single edge to nodes representing one or more additional classes (Fig. 6).

Although the core table frequently contains occurrence records linked to extension tables of images or identifications, we can also find Darwin Core archives with other structures:

- core table: event, extension table: occurrence[13]
- core table: taxon, extension tables: occurrence, specimen, distribution, reference, and description[14]
- core table: organism, extension tables: still image and identification[15]



|  | A | B | C |  |
|---|---|---|---|---|
| 1 | id | catalognumber | scientificname | genus |
| 37 | SIBA:Molluscs:131 | 131 | Deroceras (Deroceras) reticulatum (O.F. Müller, 1774) | Deroce |
| 3[illegible] | [illegible] | 132 | Deroceras (Deroceras) reticulatum (O.F. Müller, 1774) | Deroce |
| [illegible] | SIBA:Molluscs:134 | 134 | Euomphalia strigella (Draparnaud, 1801) | Euomp |
| [illegible] | [illegible] | 135 | Euomphalia strigella (Draparnaud, 1801) | Euomp |

Core table

|  | A | B |
|---|---|---|
| 1 | foreign_key | image_name |
| 2 | SIBA:Molluscs:134 | 001.jpg |
| 3 | SIBA:Molluscs:134 | 002.jpg |
| 4 | SIBA:Molluscs:134 | 003.jpg |
| 5 | SIBA:Molluscs:134 | 004.jpg |
| 6 | SIBA:Molluscs:134 | 005.jpg |

Extension table

**Figure 6.** Graph modeling the "star schema" pattern of Darwin Core core and extension tables.

- core table: preserved specimen, extension tables: PCR amplification, loan, material sample, permit, and preparation[16]

There are other more complex examples of datasets where it is not possible to use the Darwin Core Archive format because there is no single, central resource node of one type that can be related to all related resource nodes via a single edge. A more complicated graph model than the "star schema" model would be required to convert such datasets to RDF[17].

If TDWG's goal is the same as it was in 2007 (to enable data integration among providers of heterogeneous datasets), then development of a graph model complex enough to handle the kinds of examples listed above is a necessary step to follow the Darwin Core RDF Guide. There have been attempts to construct more complex graph models to express biodiversity data as RDF, such as taxonconcept.org[18], Darwin-SW[19], and Filtered Push's Darwin FP[20]. However, in order for efforts such as these to succeed, there needs to be an organized effort to involve all interested potential stakeholders and to develop a list of use cases to be satisfied by the graph model. The new TDWG Vocabulary Management specification[21] lays out a formal process for developing “vocabulary enhancements” to vocabulary standards such as Darwin Core. That process could be used to create a consensus graph model sufficiently complex to model the heterogeneous datasets in our community in a way that would allow them to be merged into a single, easily queried graph. As with the problem of lack of consensus IRIs, this is fundamentally a social challenge, not a technological one.

## 4. References

[1] Hyam R (2007) TDWG Technical Roadmap 2007. Biodiversity Information Standards Technical Architecture Group.

[2] De Giovanni R, Döring M, Güntsch A, Vieglais D, Hobern D, de la Torre J, Wieczorek J, Robert G, Hyam R, Blum S, Perry S (2010) Access Protocol for Information Retrieval (TAPIR), Version 1.0. Biodiversity Information Standards (TDWG). http://www.tdwg.org/standards/449

[3] Object Management Group (2004) Object Management Group. 2004. Life Sciences Identifiers Specification: OMG Adopted Specification. http://www.omg.org/cgi-bin/doc?dtc/04-05-01

[4] Darwin Core Task Group (2009) Darwin Core. Biodiversity Information Standards (TDWG). http://rs.tdwg.org/dwc/ (accessed on 2017-03-08).

[5] Robertson T, Döring M, Wieczorek J, De Giovanni R, Vieglais D (2009) Darwin Core Text Guide. Biodiversity Information Standards (TDWG).

[6] http://linkeddata.org/

[7] Baskauf SJ, Wieczorek J, Deck J, Webb C, Morris PJ, Schildhauer M (2015) Darwin Core RDF Guide. Biodiversity Information Standards (TDWG). http://rs.tdwg.org/dwc/terms/guides/rdf/index.htm

[8] RDF-in-HTML Task Force (2010) CURIE Syntax 1.0. World Wide Web Consortium (W3C). https://www.w3.org/TR/curie/

[9] Wieczorek J, Döring M, De Giovanni R, Robertson T, Vieglais D (2009) Darwin Core Terms: A quick reference guide. Bio-

diversity Information Standards (TDWG). http://rs.tdwg.org/dwc/terms/

[10] http://arctosdb.org/

[11] http://cetaf.org/cetaf-stable-identifiers

[12] Remsen D, Braak K, Döring M, Roberson T (2010) Darwin Core Archives - How-to Guide, version 1. Global Biodiversity Information Facility (GBIF). http://www.gbif.jp/v2/pdf/gbif_dwc-a_how_to_guide_en_v1.pdf

[13] Groom QJ, Durkin JL, O'Reilly J, Mclay A, Richards AJ, Angel J, Horsley A, Rogers M, Young G (2015) A benchmark survey of the common plants of South Northumberland and Durham, United Kingdom. Biodivers Data Journal 3:e7318. http://dx.doi.org/10.3897/BDJ.3.e7318 Dataset: Botanical Garden Meise: A common plants survey of vascular plants in South Northumberland and Durham, United Kingdom. Accessed via http://www.gbif.org/dataset/5d784d06-fa1d-4f00-8cdc-663d04d26061 on 2016-10-26.

[14] Agricultural Research Council (2016) Catalogue of Afrotropical Bees. http://doi.org/10.15468/u9ezbh Accessed via http://www.gbif.org/dataset/da38f103-4410-43d1-b716-ea6b1b92bbac on 2016-10-26.

[15] Bioimages (2016) http://bioimages.vanderbilt.edu/. 2016-05-07 release http://dx.doi.org/10.5281/zenodo.51121. Data as a Darwin Core archive in GBIF: http://doi.org/10.15468/jib4rt published 2014-07-14.

[16] GGBN (2016) Test dataset from Smithsonian National Museum of Natural History based on Global Genome Biodiversity Network (GGBN) Data Standard. http://collections.nmnh.si.edu/ipt/archive.do?r=nmnh_materialsample_test (accessed 2016-11-06).

[17] Baskauf SJ (2016) Guid-O-Matic meets the DwC-A RDF octopuses. http://baskauf.blogspot.com/2016/11/guid-o-matic-meets-dwc-rdf-octopus.html

[18] https://web.archive.org/web/20160413214827/http://www.taxonconcept.org/

[19] Baskauf SJ, Webb CO (2016) Darwin-SW: Darwin Core-based terms for expressing biodiversity data as RDF. Semantic Web Journal 7:629-243. http://dx.doi.org/10.3233/SW-150203

[20] Morris RA, Dou L, Hanken J, Kelly M, Lowery DB, Ludäscher B, Macklin JA, Morris PJ (2013) Semantic Annotation of Mutable Data. PLoS ONE 8:e76093. http://dx.doi.org/10.1371/journal.pone.0076093

[21] Vocabulary Maintenance Specification Task Group. 2017. Vocabulary Maintenance Specification. Biodiversity Information Standards (TDWG). http://www.tdwg.org/standards/642