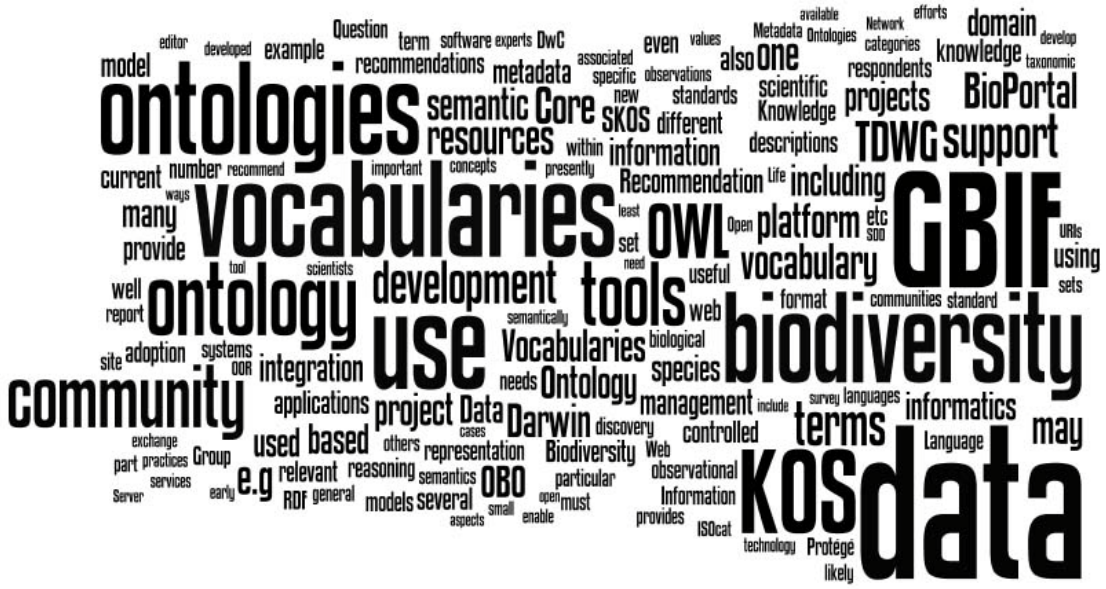
Recommendations for the Use of Knowledge Organisation Systems by GBIF

Version 1.0

Terry Catapano[1], Donald Hobern[2], Hilmar Lapp[3], Robert A. Morris[4], Norman Morrison[5]
Natasha Noy[6], Mark Schildhauer[7], David Thau[8]



February 2011

---

[1] Librarian, Columbia University Libraries and Vice President, Plazi
[2] Director, Atlas of Living Australia and TDWG Chair
[3] Asst. Director for Informatics, National Evolutionary Synthesis Center (NESCent)
[4] Convener; Professor Emeritus of Computer Science, Univ. of Massachusetts/Boston and Informatics Associate, Harvard University Herbaria
[5] Ontologies and Data Standards Manager, Natural Environment Research Council, Environmental Bioinformatics Centre (NEBC) & The University of Manchester, UK
[6] Senior Research Scientist, Stanford Center for Biomedical Informatics Research(BMIR), Stanford University
[7] Director of Computing, National Center for Ecological Analysis and Synthesis(NCEAS)
[8] Developer Advocate, Google

Suggested citation:

GBIF (2011). Recommendations for the Use of Knowledge Organisation Systems by GBIF. Released on 04 Feb 2011. Authors: Terry Catapano, Donald Hobern, Hilmar Lapp, Robert A. Morris, Norman Morrison, Natasha Noy, Mark Schildhauer, David Thau. Copenhagen: Global Biodiversity Information Facility, 49 pp., accessible online at http://links.gbif.org/gbif_kos_whitepaper_v1.pdf.

Persistent URI: http://links.gbif.org/gbif_kos_whitepaper_v1.pdf

Language: English

Document Control:

Version	Description	Date of release	Author(s)
	Report as submitted by KOS Task Group	20/12/2010	Terry Catapano, Donald Hobern, Hilmar Lapp, Robert A. Morris, Norman Morrison, Natasha Noy, Mark Schildhauer, David Thau
1.0	Copy editing	04/02/2011	Éamonn Ó Tuama

Preface

It was a great pleasure, and a tremendous learning experience, to work with the team I put together in response to the call for a White Paper to offer GBIF advice on the application of Knowledge Organisation Systems. Given a desire to keep the group small, I think we nevertheless had the breadth and depth to do the job, though by no means were uniquely qualified to do so. We're grateful to colleagues who took the time to publicly comment on an early draft, and hope the community will continue to offer GBIF their opinions on omissions of important resources, on the recommendations we make, and on the strength or weaknesses of the arguments supporting them. Also thanks to GBIF Program Officer Éamonn Ó Tuama for insights into some of GBIF's relevant current activities.

Many readers will know the cost of drafting by a committee of experts: it is that they are all very busy, working under many deadlines besides those of producing such a report as this. Consequently, not everyone could always follow the evolution of my drafting on the same schedule as I produced it, even when I was simply inserting text of theirs. The result is that any remaining lack of clarity, errors of omission or fact, or citations of obsolete references must be laid at my feet.

There was remarkably little contention about the recommendations. What contention there was, centred on the pace at which GBIF should move toward broad, deep, biodiversity ontologies. Some argued that current research on biodiversity ontology-driven applications should be where GBIF immediately sets its sights. Others argued that domain scientists would not accept the current tools and instead should be immediately enlisted to assist with lower hanging fruit, especially focused on the ontology principles that are easy to explain, such as hierarchical concept vocabularies. No doubt some of the group will even argue that I haven't framed the debate properly. This debate probably cannot be resolved on *any* deadline, much less the one before us. So in the section "Next Steps", I try to summarise that the way forward is for GBIF to simultaneously act on *both* positions, the first in close coordination with the informatics community that is providing data *now,* while also working in collaboration with the research projects dedicated to adding farther-reaching semantic value to biodiversity data. Once again, deadlines have left little time for spirited debate on my framing, so once again any lack of clarity, or failure of the argument, falls on me.

Robert A. Morris, Convener

Boston, December 2010

Contents

List of Tables

List of Figures

Executive Summary

This report responds to an Request for Proposal[9] by GBIF for recommendations that will inform its planning for deployment, provision and support of knowledge organisation tools for the management, service, and use of biodiversity data by the GBIF Secretariat and membership.

With *Data* as its foundation, *Knowledge* sits near the top of the *Data-Information-Knowledge-Wisdom* pyramid. Organising *Wisdom* is of course the domain of the consumers of the *Knowledge* that GBIF can enable, but GBIF and the communities it serves, and those with which it collaborates, have long been active in approaching the lower two blocks of the pyramid. Such efforts quickly recognise the fact that users of the information systems may come to radically different conclusions from the same *Data* and *Information*. The real problem is those users may have no way to recognise that such inconsistencies have happened, much less why. As in many disciplines, early approaches to this conundrum focused on community agreement about the definition and use of controlled vocabularies for the scientific concepts in use, and for context concepts, e.g., for specification of times and places, of individuals and organisations, of data gathering methodology, etc. In an era of extensive internet connectivity providing access to petabytes of highly heterogeneous scientific, social, and organisational data, controlled vocabularies remain central, but insufficient. All aspects of the *Knowledge* layer need technological assistance, even to the level of controlled vocabularies for describing controlled vocabularies themselves. The computer and social systems that accomplish this are the subjects of the discipline known as Knowledge Organisation Systems (KOS). KOS encompasses various semantic approaches to labelling and interpreting digital data—using controlled vocabularies, metadata specifications, gazetteers, thesauri, and ontologies, etc., along with standards-based tools, to provide advanced technology services that are compatible with the Semantic Web. In the area of biodiversity informatics, these approaches might provide powerful and efficient ways of representing and exchanging biological taxonomies and other types of biodiversity-relevant information over the Internet. Hodge (2007) is an often cited survey of KOS. The needs of the biodiversity data community for KOS, the current state of biodiversity KOS, and recommendations to GBIF for its involvement with KOS are the subject of this white paper.

[9] http://www.gbif.org/communications/news-and-events/showsingle/article/request-for-proposals-for-a-position-paper-on-vocabularies/

1 Background and Context

Users of scientific information systems often understand the output of those systems in very different ways, even in response to the same query. The problem is worse when data are aggregated from different databases by mapping each to a common federation schema, because there is typically no guarantee that the mappings are consistent with one another. While such inconsistencies are perhaps inevitable given the heterogeneity of data and databases, end-users often have no way to recognise them, much less what caused them. Early approaches to this issue focused on community agreement about the definition and use of controlled vocabularies for the scientific concepts in use, and for context concepts, e.g., for specification of times and places, of individuals and organisations, of data gathering methodology, etc. In an era of fast online access to petabytes of highly heterogeneous scientific, social, and organisational data, controlled vocabularies remain central, but insufficient. All aspects of the creation of knowledge from data need technological assistance, even to the level of controlled vocabularies for describing controlled vocabularies themselves. The computer and social systems that accomplish this are the subjects of the discipline known as Knowledge Organisation Systems (KOS). KOS encompasses various semantic approaches to labelling and interpreting digital data—using controlled vocabularies, metadata specifications, gazetteers, thesauri, and ontologies, etc., along with standards-based tools, to provide advanced technology services that are compatible with the Semantic Web. In the area of biodiversity informatics, these approaches might provide powerful and efficient ways of representing and exchanging biological taxonomies and other types of biodiversity-relevant information over the Internet. Hodge (2007) is an often cited survey of KOS. The needs of the biodiversity data community for KOS, the current state of biodiversity KOS, and recommendations to GBIF for its involvement with KOS are the subject of this white paper.

The group of authors is geographically diverse (one Australian, one European, and six from the U.S.A.), represent a broad spectrum of informatics expertise relevant to KOS, and are involved with the production of KOS tools in a variety of biodiversity-science related domains, including environmental, ecological, and evolutionary science. Nonetheless, relevant KOS projects may have been inadvertently omitted from this report, and we thank GBIF Program Officer Éamonn Ó Tuama and several public reviewers for pointing out some of those earlier. The public is invited to continue providing comments and input[10] during GBIF's action on this report, including discussion of our recommendations.

1.1 Context from the GBIF Report on Vocabularies for Biodiversity.

This Report was commissioned under a Request For Proposal[11], which, besides its charges, set as its context a previous paper, the GBIF Report on Vocabularies for Biodiversity (GRVB)[12]. To set the context for our more extensive report and more specific recommendations, we here respond to the six Recommendations. Specific KOS tools, other resources, and detailed explicit recommendations are described later in this work.

GRVB Recommendation 1: *To ensure rapid convergence on agreed terminology for biodiversity informatics, GBIF should promote the practice of developing flat vocabularies (concepts and their definitions) as an independent activity from modelling relationships between concepts.*

[10] http://bit.ly/GBIFKOS_Comments

[11] http://www.gbif.org/communications/news-and-events/showsingle/article/request-for-proposals-for-a-position-paper-on-vocabularies/

[12] http://imsgbif.gbif.org/CMS/DMS_.php?ID=1057

We recommend this direction and believe that it must be accompanied by the development of materials and outreach that make it clear that the principal benefit of flat vocabularies is that they promote data exchange. Textual concept definitions can describe the intended semantics, but software tools cannot use those descriptions to understand, let alone enforce them. This is true even for simple semantics such as a requirement that two data fields both either be present or absent. For example, the Darwin Core cannot express in a machine-interpretable way that a record with a value for decimalLatitude but none for decimalLongitude cannot be meaningfully compared to others by geolocation. Nevertheless, documentation for such vocabularies can specify *intended* relations among the terms, and flat vocabularies should be developed so that they are reusable as a terminological foundation for semantically richer vocabularies or ontologies.

GRVB Recommendation 2: *Evaluate various platforms that provide both a social and technical mechanism for vocabulary and ontology development including i) the TDWG ontologies site, ii) the Google Docs / TDWG wiki system used for Darwin Core, iii) the OBO Foundry system, iv) the ISOcat system, v) the GBIF vocabularies site, and vi) Collaborative Protégé; provide recommendations on the best system/combination of systems (henceforth referred to as vocabulary servers) for managing the development, publishing and maintenance of vocabularies.*

Here we briefly discuss our evaluation of the mentioned platforms and others. In a later section we offer more specific details in some cases.

i) The TDWG ontologies site is limited and purpose-built. It is presently mainly a collection of static OWL files with a user-friendly HTML presentation. It is likely that almost any solution to ontology life cycle management can support that presentation if required, since it is based on CSS. See TDWG Vocabularies for further details.

ii) Google Docs' main virtue is that it is a collaborative text editor allowing multiple simultaneous editors. Beyond that it lacks practically any specific support for ontology life-cycle management, ontology navigation, or presentation, and we therefore cannot recommend it as the technology platform for vocabulary or ontology development. In addition, some countries block it, and any necessary software development with the Google Docs API runs the risk of becoming obsolete if Google decides to discontinue the product.

iii)We strongly recommend consideration of the OBO Foundry governing principles. However, the OBO Foundry is not a technology or software platform, and the functionality of the tools it uses can mostly be found in other platforms as well. Hence, consideration or adoption of the (social) governing principles need not be tied to adopting the same technology platform.

iv) ISOcat has a noteworthy user-friendly UI, which should be consulted for user-interface design decisions of any future technology development, whether built on ISOcat or not. That notwithstanding, as a candidate technology platform for adoption by GBIF, ISOcat raises enough concerns that we cannot recommend it at this time for serious investment of resources by GBIF. Specifically, ISOcat has a very small uptake, by the single community (Linguists) for whom it was designed, and which is rather unrelated to biodiversity or biological science. It is part of CLARIN, an EU project which is just now starting its planning stages, and thus far from reaching maturity, and so ISOcat may be dependent on CLARIN for its sustainability story.

v) Continued maintenance of the GBIF Vocabulary Server is low risk, as it is based on the EU ScratchPads project (which itself is based on Drupal). Hence, the sustainability of the technology itself is likely high. However, it is not yet clear how much uptake there will be of ScratchPads in general and their use for vocabulary development in particular, and its feature support for the vocabulary development life-cycle lags far behind other software platforms. Thus, we recommend that further investment in a Scratchpad-based GBIF

vocabulary server always be based on clearly stated use-cases that demonstrate its value over full-featured systems specifically developed for managing ontologies and vocabularies (such as the BioPortal).

vi) Collaborative Protégé is a highly useful tool, though just emerging in its OWL2 form. The fact that it can handle both OBO and OWL is a large advantage, since these two formats have the largest uptake in the biological sciences. It works well with the NCBO BioPortal platform, which we elsewhere recommend together with Collaborative Protégé.

vii) CmapTools is a free software suite for collaborative construction, sharing and publishing of knowledge models represented as concept maps. Its web site claims it is used worldwide by millions of users, and is available in over 15 languages. Concept maps are a gentle way to introduce general principles of knowledge representation, and the software is very easy to use, at least for individuals (some of the emphasis is on collaboration).

In general, since fully integrated lifecycle tools are the goals of many projects, some in early stages of development or integration, it is important that GBIF keep an ongoing focus on increased convergence and interoperability of tools in this rapidly changing landscape. Of particular importance is to remain aware of APIs and specifications dedicated to tool integration. Examples include the mature OWL API[13] and the emergent Open Ontology Repository (OOR) [14] Initiative high level requirements. We discuss both elsewhere.

Time prevented us from evaluating the success and applicability of specific governance arrangements, and that task remains to be done. However, we know that ontologies have a much greater likelihood to be agreed upon and become adopted by communities of practice if the respective domain experts have been engaged in their development early on. Furthermore, ontologies capturing domains of scientific knowledge cannot be static. To stay relevant and useful, they must be able to evolve continually as scientific knowledge changes in the light of new research and new insights. Keeping domain experts continually involved and engaged in this evolution is at least as, and perhaps even more, important than for the initial development.

This need for managing and engaging the community of domain experts pertains to biodiversity science as much as to other domains of biology, and has indeed been recognised for a long time by ontology building initiatives in the biomedical and molecular biology domains. For example, the ontologies in and around the OBO Foundry have established means that track the rationale for proposed term changes and additions, and the status of the request. In the past this has been implemented using the bug tracking features of source code repositories (specifically, SourceForge), which in theory is also being used by the Darwin Core team (using Google Code's bug tracker). In practice, years of experience within the OBO community with this system have revealed a variety of issues, including difficulty of easily and reproducibly identifying a tracker request; transparently connecting comments on the tracker item to pertinent mailing list traffic and vice versa; difficulty of recording comment threads as provenance information that is permanently linked to the term; and usability issues that impede the continual engagement of the community of relevant experts. In recognition of these problems, a number of commenting, notification, and status change alert features have recently been added to the BioPortal software, and are undergoing continual enhancement.

While debates of the correct concept hierarchy will not apply to flat vocabularies, all other properties of terms (such as meaning, definition text, applicability, etc) still do, and thus there is no fundamental difference between ontologies (or thesauri) and vocabularies as far as the needs for capturing and tracking the input from experts are concerned.

[13] http://sourceforge.net/projects/owlapi/

[14] http://ontolog.cim3.net/cgi-bin/wiki.pl?OpenOntologyRepository

Therefore, a technology solution for managing vocabularies, flat or not, will encounter similar needs and issues as have been experienced in other domain science-rooted ontology building communities. It would hence seem prudent for GBIF to tap into the software platforms developed for those communities, rather than starting its own, disconnected effort based on a system that has not been vetted, or informed, by its ability to keep domain scientists looped into ontology development.

GRVB Recommendation 3: *Test a proposed TDWG three-layered approach for ontology development: layer i): like DwC – well defined human readable terms that are not constrained as to domain and range; layer ii): logical layer with OWL for models and XML schema for exchange between applications; layer iii): project docs including applicability statements which define which items and technologies in layer ii to use. Based on recommendation 2, determine the platform that should be adopted for ontology modelling.*

We endorse the Recommendation with two reservations: First, we understand "XML schema" in this case to mean "RDF in XML". While the current tools for exchanging data described by OWL ontology in XML may be in wide use, nothing in the OWL layer should foreclose exchange in other RDF forms, particularly as other forms become popular for exchange. Second, some aspects of layer iii) are required for layer ii), because they should be part of the design process for ontologies. An important case is support for use cases and competency questions[15], which we urge be a specific part of any implementation of the logical layer. A critical service for all layers is the issuance of persistent Globally Unique Identifiers for terms and relations.

Project documentation, especially as to applicability, should also support determination of the utility and applicability of vocabularies and ontologies other than GBIF's. The community mechanisms for discussing and adopting applicability statements should closely follow those for vocabularies and ontologies, and in particular be a resource for GBIF to marshal collaboration with similar efforts in other groups.

GBIF should evaluate the plans and activity of the Open Ontology Repository (OOR) Initiative as to suitability for deposit of GBIF-developed ontologies, and consider adopting the OOR High Level Requirements as its own repository requirements. In any case, GBIF should actively participate in the OOR community. We discuss OOR in more detail elsewhere.

GRVB Recommendation 4: *GBIF should promote use of technologies and standards that support multi-lingual vocabularies/thesauri/ontologies. GBIF/TDWG should adopt SKOS as a mechanism for expressing and sharing multi-lingual vocabularies.*

We strongly endorse the general point. We would place some qualifications on the adoption of SKOS. It has wide acceptance and is now a W3C Recommendation. However, by its history and design, its core for mapping and expressing semantic relations between domain concepts is necessarily coarse. Without extension, SKOS is mainly useful for thesauri, or thesaurus-like exploration of more structured vocabularies. By contrast, for example, the OBO Relation Ontology defines some term relationships that provide biological semantics, such as part hierarchies. The current SKOS has a minor subset that is an OWL DL ontology. Restricting to this subset would: enable the use of OWL tools; ease controlled, biologically relevant extension of SKOS supporting tractable reasoning on the

[15] Competency questions are an ontology design paradigm analogous to use case scenarios. They are formulated as queries, either formal or informal, against which the ontologies and a knowledge base may be tested for efficacy and scope. Using the questions, the designer can identify the concepts and their relationships needed for the ontology. A brief overview with examples is at http://marinemetadata.org/references/competencyquestionsoverview .

descriptions of vocabularies; and possibly decrease the costs of future migration to more "biologically relevant" descriptions of vocabularies. Because of this, we recommend that SKOS should be a *minimal* mechanism for expressing descriptions of biodiversity vocabularies no matter how well structured, and more expressive descriptions should be the ultimate goal. The OWL 1 DL version of SKOS should be mandated for SKOS descriptions, and GBIF should follow SKOS developments as to OWL 2. We also note that W3C designed SKOS as an extensible language[16] and thus recommend that if additional inter-vocabulary relations are required, they are modelled as SKOS extensions. Finally, we note that there are tools emerging dedicated explicitly to producing SKOS-described thesauri, and by their simplicity might be useful in workshops aimed at domain scientists with no experience in ontology engineering.

Substantial attention has been given to multilingual thesauri, e.g., for SKOS[17] but deeper multilingual semantic technologies have been largely focused on static web pages for discovery, and natural language generation for publishing. Multilingual reasoning issues remain a research area (e.g., Buitelaar et al., 2010). The current SKOS provides explicit support for signifying alternative labels in different languages, and for providing mappings between ontologies in different languages. In general, RDF supports specifying the language of an *rdfs:label* [18]with the *language tag* on RDF literals. Deeper reasoning than language comparison must rely on label metadata properties like those of SKOS. For example, the Semantic Web Rule Language (SWRL)[19] provides similar support for rule-based reasoning on OWL ontologies.

GRVB Recommendation 5: *Any vocabulary servers adopted in recommendation 2 should include a mechanism for providing persistent identifiers for both the vocabulary and its constituent terms.*

We strongly endorse this and would replace "should" with "must", and would extend the principle to all three layers of Recommendation 2. We also strongly recommend that a robust mechanism for term versioning be adopted and note that this is likely to require a mechanism for, and commitment to, maintaining maps that align vocabularies over time. The provision and versioning mechanisms should be supported both in tools for human use and frameworks for machine access.

GRVB Recommendation 6: *The GBIF GBRDS / Metadata catalogue should accommodate entries on biodiversity related vocabularies in order to aid discovery and re-use. A minimal schema for describing vocabularies should include: title, description, standard reference, responsible party/organisation/web site, persistent identifier, access URI."*

We endorse this recommendation but add that any such description must be consistent with any ontology registry entries GBIF might make in its own, or community registries. This implies that a standard mapping must be maintained between GBRDS description terms and terms used in any other descriptions, such as those that may be closely coupled with the vocabulary management tools.

[16] http://www.w3.org/TR/skos-reference/#xl

[17] http://www.w3c.rl.ac.uk/SWAD/deliverables/8.3.html

[18] http://www.w3.org/TR/rdf-schema/#ch_label

[19] http://www.w3.org/Submission/SWRL/

2 Findings

2.1 Needs for KOS

Two approaches contributed to the assessment of the needs for the deployment of Knowledge Organisation Systems by GBIF. The first was a brief community survey[20]. Its analysis follows next. Following that is a synthesis of the authors' own view of the needs.

2.1.1 Identification of needs as perceived by community (analysis of survey)

The survey was exploratory, not hypothesis driven. It was intended to elicit community perceptions of the uses for KOS, the impediments to KOS deployment, and the current level of KOS awareness and expertise of the respondents. As of December 11, 2010, the survey had responses from 99 distinct individuals. 10 questions were some form of multiple choice, one was open ended, and one asked for voluntary contact address for follow-up. For no question was an answer obligatory. 44 of the 99 volunteered an email address. At least 97 of the respondents answered 7 of the multiple choice questions, and at least 66 answered the other three. From this we conclude that the survey kept the respondents engaged.

This informal analysis discusses only those questions that we think lead immediately to recommendations. Perhaps because the self-identified roles of respondents within their organisations varied widely, some questions also seem to show no clear trends based only on informal examination. We recommend therefore that the survey be kept as an ongoing effort, if more responses become available, that correlations between roles and answers be mined for more accurate descriptions of the needs of various portions of the community.

KOS familiarity. The graph below (Figure 1) represents the Question 3 self-assessment of respondents about their familiarity with several categories of KOS concepts. If we define "familiar" as either "familiar or very familiar with one or more" examples of a given category, and in each category sum the corresponding percentages, then about 65% are familiar with Controlled Vocabularies, slightly more than 53% with Gazetteers, and slightly fewer with Ontologies and Thesauri. Perhaps remarkably, only 38% expressed familiarity with Linked Data[21], a set of linking standards presently garnering a lot of discussion in some segments of the Semantic Web community. We discuss it elsewhere in this report. Concept Maps brought little claim of familiarity (31%). In retrospect, we feel that this should not have been included in the question, since Concept Maps represent more of a tool than a category of resource.

[20] http://www.surveymonkey.com/s/GBIFKOSurvey

[21] http://linkeddata.org/

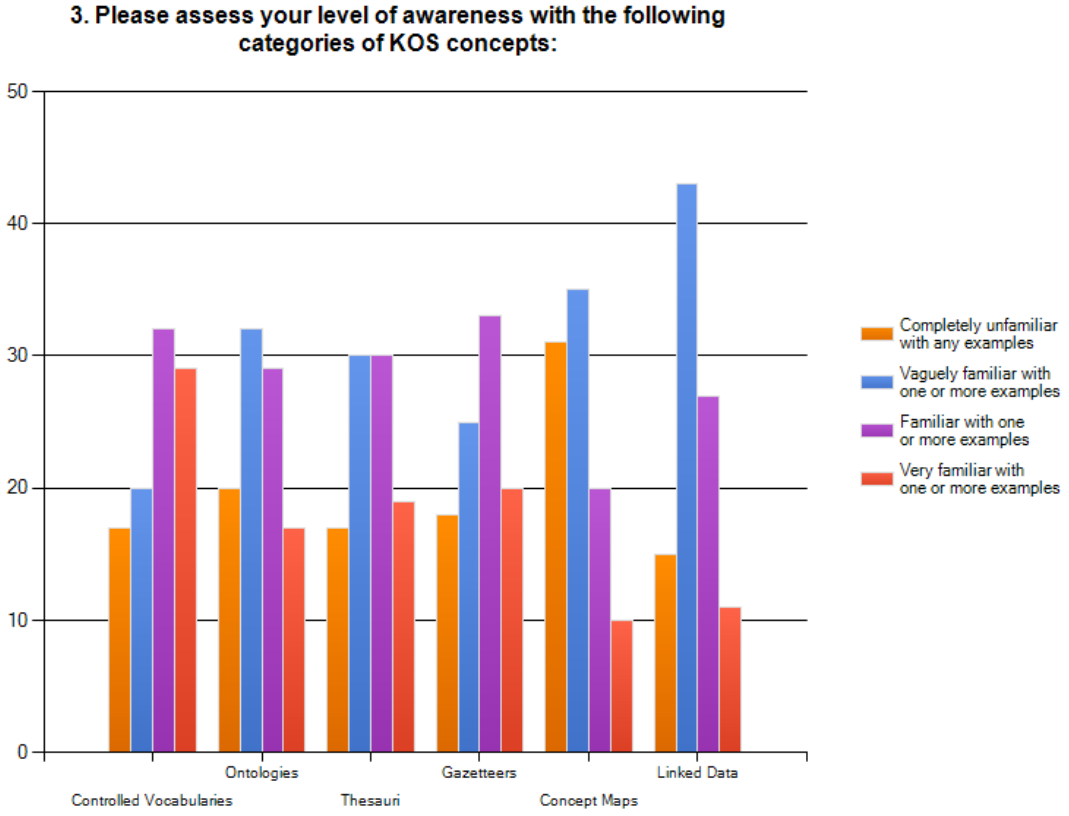


Figure 1. Survey Question 3.

Tool familiarity (Question 4). We asked how familiar respondents are with four categories of tools: Ontology editors, ontology visualisers, reasoners, data mining tools. For each of these categories, fewer than 20% of the respondents indicated that they are familiar or very familiar with at least one tool in the category. Taken together with the responses for Question 3, one explanation is that the respondents are mostly consumers, not producers, of KOS resources.

Aspirations for KOS. The data from question 6 (Figure 2), visualised below, suggests that data discovery, integration, and linking are the largest aspirations respondents have for KOS, followed closely by the reduction of ambiguity while interpreting data. It is notable that only a minority aspire to reasoning applications, although many of the other KOS uses are accomplished with the aid of reasoning. This presumably reflects the relative immaturity of KOS solutions within biodiversity informatics and general experience of the fact that even basic integration of biodiversity data presents many complexities.

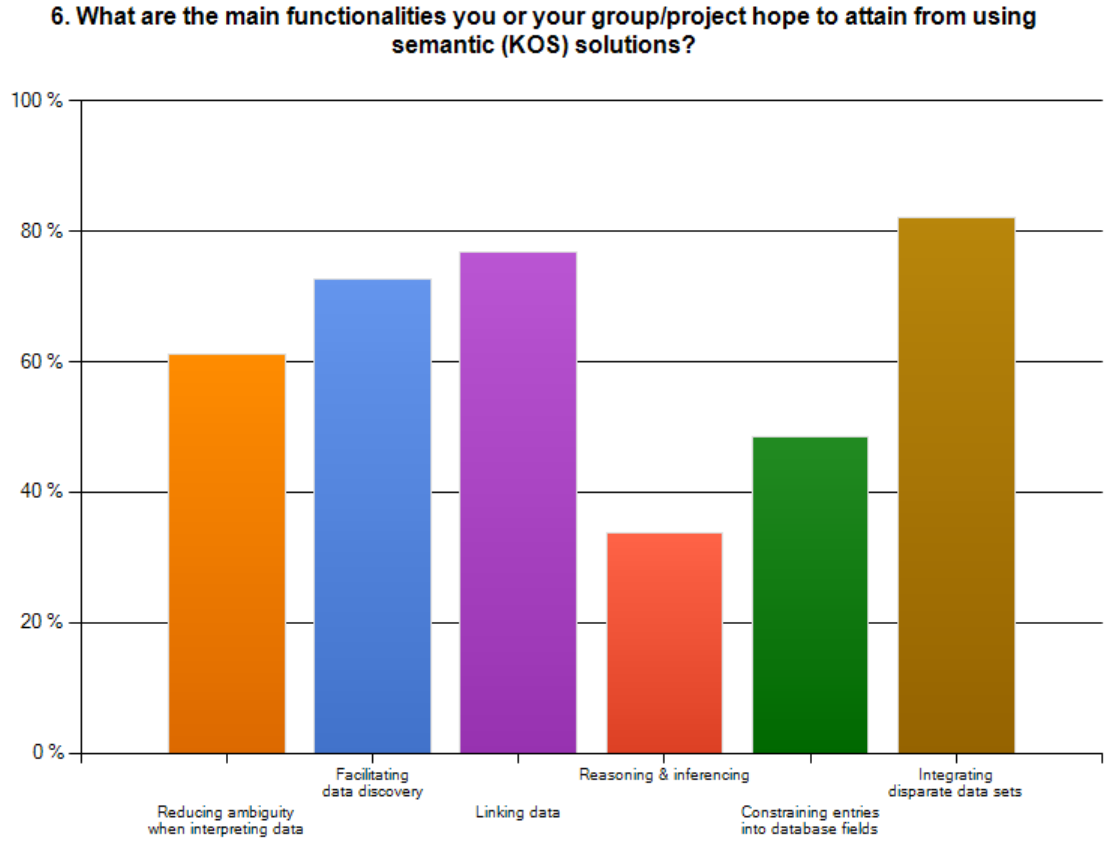


Figure 2. Survey Question 6.

What standards are you and your organisation currently considering or using for semantic solutions (Question 8). Over 30 different controlled vocabularies already in use were mentioned in the open-ended response to Question 8 of the survey. Unsurprisingly, Darwin Core is the most frequently mentioned. 85% of the respondents indicated in the multiple choice section that they used XML metadata and 59% use RDF. But only a minority of the 66 respondents indicated that they use any of the formal ontology languages listed (OWL, OWL2, OBO, SKOS) or use SPARQL queries.

Impediments to adoption. It is no surprise that the largest reported impediments to adoption (Figure 3) are insufficient funding and insufficient technical support. Our main conclusion from this is that GBIF will not face organisational impediments different from what it previously has faced in bringing leading-edge tools to its members. Socially, it is perhaps noteworthy that fewer than 20% expect resistance based on a belief that semantic approaches are not useful. In retrospect, perhaps we should also have included "Complexity of domain" as a possible impediment. It is quite possible to believe that semantic approaches could be useful in general, even if there are doubts that the quality of the data and the rigour of the KOS tools are currently up to the task.

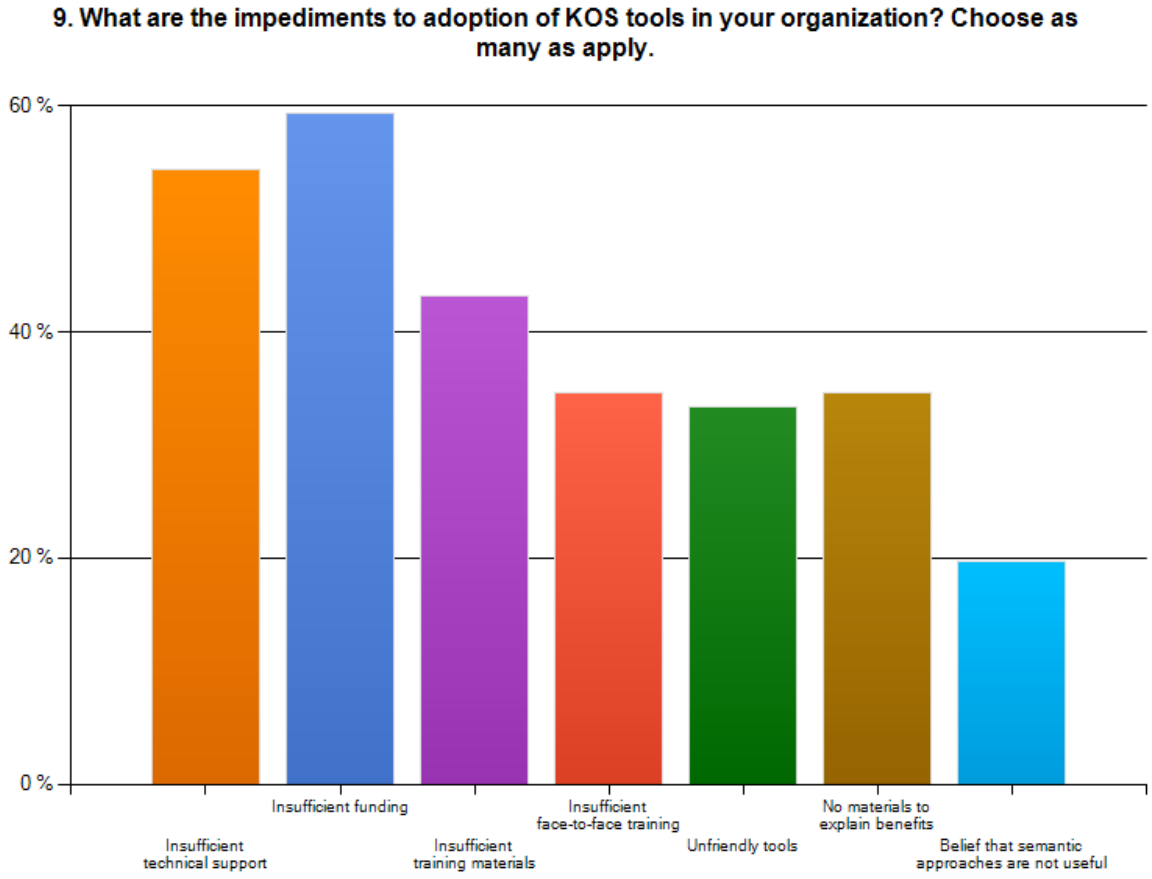


Figure 3. Survey Question 9.

Desirable attributes of technology. In Question 10, we asked respondents to rank 8 attributes of technology, shown in the left column of the table below (Table 1). If we rank these attributes based on what fraction of the respondents put them in the top 2 of their rank and also rank them by those choosing them in the top 3, only a few hypotheses seem well supported: (1) Most respondents put good documentation ahead of everything else. This will lead us to recommend that GBIF invest strongly in documentation of the KOS tools, practices, and resources it adopts or encourages. (2) Most respondents put availability of APIs, Face to face training and Expressivity at or near the bottom of their list of desiderata. One explanation for this may be that most respondents have exposure to KOS as users but not as developers. That is consistent with the analysis above of Question 4.

Table 1. Re-ranking of Survey Question 10.

Re-ranking of Question 10				
Criterion	**% choosing in top 2**	**Rank based on top 2**	**% choosing in top 3**	**Rank based on top 3**
Good documentation	76.4%%	1	89.9%	1
Good UI	58.0%	2	74.4%	2
Wide adoption	53.5%	3	71.6%	3
Programming ease	48.2%%	4	65.1%	4
Online training	38.8%	5	57.6%	5
Availability of APIs	43.5%	6	56.5%	6
Face-to-face training	26.7%	7	47.7%	8
Expressiveness	26.4%	8	54.2%	7

2.1.2 Identification of needs as perceived by experts (team and others)

There appear to be no systematic attempts to develop formal, stylised use cases, competency questions, or other goals for use of KOS in the biodiversity informatics community, against which specific solutions can be tested.

As such goals emerge, domain scientists need easy-to-use KOS tools that transparently manage the life-cycle of biologically relevant knowledge representation resources. Present KOS production tools are far too difficult for use by domain scientists. Users also face a plethora of controlled and uncontrolled vocabularies, whose variety of syntax and terms vary wildly across and within sub-disciplines. Consumers of biodiversity data need transparent discovery and integration of data, with queries answered only with relevant responses. Of course, these needs are not particular to biodiversity data producers and consumers. They apply throughout the sciences. But in Section 2.3 ("Gaps in Current Biodiversity KOS") below we identify a number of specific gaps in the current state of biodiversity KOS, and we may simplistically say: those gaps need to be filled.

Evolutionary biology and Linnaean taxonomy provide hierarchical knowledge organisation systems. However, both depend on sometimes sparse data (e.g., in the fossil record or in the rapidly increasing genomic record), on opinions of specialists as to the most appropriate choice and value of characters upon which to base taxonomies, and on opinions about which algorithms are best for selecting evolutionary trees from among several. In practice, biodiversity knowledge also suffers from the fact that identifying taxa is a non-trivial act and open to considerable debate. Taxonomic judgment on the boundaries of taxa will vary. Ability then to identify an organism to a recognised taxon concept and to provide a robust reference to that concept is equally difficult. This means that most of the data available for integration have an unreliable connection to the species or taxon to which they have been identified, even though taxonomy should be the core axis on which scientists rely for subsequent inference. Without robust solutions to these issues, many practitioners trust their experience—perhaps justifiably—more than computer-based inferences.

Finally, even for long established legacy KOS such as Darwin Core, nomenclators, term lists, collections lists, checklists, etc., there appears to be no semantically enabled discovery of these resources. Work across sub-disciplines is hampered by this, as scientists haphazardly find resources which may or may not be the best fit for their purpose. For example, a field biologist made aware of ITIS might never become aware of its relationship to the Catalogue of Life. (That relationship is mentioned on the ITIS front page, but not very prominently, and, at the time of this writing, it is not featured on the logo-based page of partners). The ability to exploit some simple relationships of resources or projects to one another (e.g. partnersWith, isComponentOf, usesAsNameAuthority, etc.) could dramatically reduce duplication and increase awareness of resources.

2.2 Current state of biodiversity KOS

Prior to the development of recent web-based biodiversity data applications such as the GBIF Portal[22], many standard data sets were developed to support consistent databasing of botanical and zoological information within and between different institutions. Many of these early resources are documented by the Berlin Botanical Garden[23]. Some of these are in current use by biodiversity informatics projects, but very few are available as machine-accessible vocabularies. Normalizing existing data from different sources to use these vocabularies would be a significant effort, but tools might be developed to identify likely matches, e.g., for journal titles or natural history collection, which, for example tend to have good authority lists.

Below we list a number of KOS resources in use in biodiversity information systems. We do so in several categories: 1) Vocabulary formats and languages; 2) Vocabularies and ontologies; 3) Data providers; 4) Software platforms, projects, and practices; 5) Ontology life cycle tools. These categories necessarily overlap, and some entries will have aspects of several categories. All of them are meant to be illustrative, and not all necessarily play direct roles in our recommendations. Certainly none are exhaustive. Particularly the "Data providers" category represents but a very small sample, some central and some not.

2.2.1 Vocabulary formats and languages

The OWL Web Ontology Language[24] **("OWL 1")** and its recently recommended enhancement **OWL 2**[25] are web ontology languages now established as Recommendations

[22] http://data.gbif.org/welcome.htm

[23] http://www.bgbm.org/TDWG/acc/Referenc.htm

[24] http://www.w3.org/TR/owl-features/

of the Worldwide Web Consortium. The OWL family of languages arises from a twenty year history[26] of computer ontology representation and use, with particular attention to the use of formal logic to add meaning to data.

OWL 1 and OWL 2 each have a set of recommended standard syntaxes and semantics. One syntax is based on RDF, the W3C Resource Description Framework[27], which itself has a representation in XML, making it exchangeable with a wide variety of existing tools. OWL 2 provides a number of profiles[28] that address different uses of ontologies. Together with several new features, these provide not only greater expressivity than OWL 1, but also the ability to circumvent tractable computing issues that arose when attempting to add semantic modelling to data held in relational databases. Several widely used subsets of OWL 1 are also consistent with OWL 2, and there are tools that aid in determining whether OWL 1 ontology meets the OWL 2 specifications in one or another profile. An important new feature of OWL 2 is support for limited independent reasoning on internal annotations of the terms in OWL 2 ontology. Of particular note about annotations is that OWL 2 provides the ability to identify versions of an ontology and to assert that a term is or is not compatible with a previous version.

The **Entity-Quality (EQ) model** (Mungall et al., 2010), an ontology-based approach enabling formal logic reasoning, is being applied to the biodiversity of phenotypes as documented in the comparative systematics and taxonomic literature in the form of natural language. The formalism was originally developed within the model organism community to allow integrative analysis of mutant phenotypes across species (see Washington et al. 2009, Mungall et al. 2010). It decomposes phenotype descriptions into three main components: a phenotypic quality (Q), such as an 'elongated' shape or a ‘red’ colour; the entity that is its bearer (E), such as an anatomical structure; and the organismal entity that exhibits the phenotype, for example an observed individual, members of a taxon, or the carriers of an allele. Phenotypes in EQ-format consist of terms from requisite ontologies for each component, and relations that render them formal logic expressions. Phenotypes expressed in this way are interoperable and can thus be integrated across sources. Furthermore, machine reasoners can use the subsumption and other hierarchies of the ontologies from which the terms are drawn to infer facts that are implied, but not asserted, among a set of EQ phenotypes.

The **OBO format** ontologies are based on the EQ model, and typically support type (“is_a”) hierarchies and meriological (“part_of”) hierarchies. These being common in biological domains, OBO format ontologies have recently been demonstrated to be particularly useful for semantic data discovery and integration (e.g., Mungall et al. 2010)

There are a number of **other ontology languages**[29]**,** many arising from Artificial Intelligence. One of particular legacy importance is the class of **Frame-based**[30] languages, in which at least one ontology[31] of biodiversity interest was originally developed. In general, ontologies specified in one language may be expressible several ways in another,

---

[25] http://www.w3.org/TR/owl2-overview/

[26] http://en.wikipedia.org/wiki/Web_Ontology_Language#History

[27] http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

[28] http://www.w3.org/TR/owl2-overview/#Profiles

[29] http://en.wikipedia.org/wiki/Ontology_language

[30] http://www.obitko.com/tutorials/ontologies-semantic-web/frame-based-models.html

[31] The Foundational Model of Anatomy (FMA), discussed in Sec. 9.2.2.

so it is important to know how the transformation was done. A similar issue arises in representing OBO format ontologies in OWL[32]

2.2.2 Vocabularies and ontologies

The **Darwin Core (DwC)** and **ABCD**[33] are the de-facto standards for biodiversity occurrence data. The new Darwin Core[34], adopted by TDWG in 2009 is a representation-free vocabulary in the style of Dublin Core[35] (and using some of it). It has an RDF representation, but that is largely free of formal semantics addressing its biological concerns. Recently, the TDWG-Content mailing list has had extensive discussion surrounding moves toward a DwC RDF representation that will facilitate reasoning. ABCD is a significantly more complex schema for representing and exchanging information on materials held in natural history collections. Its structure as a complex nested XML schema, with some semantics implied by the document structure, makes it less flexible for new applications than DwC. It should nevertheless be considered a significant and inclusive attempt to model collections data for interchange.

The term vocabulary of DwC[36] has an RDF representation that is minimalist in style, somewhat modelled on the most recent design of the Dublin Core. There are several technical pathways to making it, or something like it, an OWL 2 ontology and any such efforts will require community agreement about the knowledge representation use cases which such ontologies must enable. Thus, there remains a question about whether an OWL representation matters for DwC, but see the discussion below of Linked Data.

Documentation for many of the DwC terms, such as basisOfRecord, lifeStage and sex, includes recommendations or suggestions of typical free text values which may be used for these terms. These recommendations are not mandated, although the GBIF Vocabularies Server provides URIs and a lookup service[37] for the recommended values. This is a reflection of the use of DwC as a flexible transport mechanism for species occurrence data from widely differing sources, but there is consequently little possibility for data consumers to rely on standard values as a basis for stringent data integration or reasoning. Some of the GBIF vocabularies (see below) and similar vocabularies from other sources could be candidates for use in this context, but the TDWG Darwin Core Task Group should then document how data providers should serve controlled terms including URIs in place of, or in combination with, free-text values. The existing availability in DwC of separate verbatim and interpreted terms (e.g. verbatimEventDate alongside eventDate and verbatimLatitude alongside decimalLatitude) might be a model for this. However, since the existing DwC terms such as basisOfRecord are already extensively used for free format values, the introduction of new terms such as basisOfRecordURI may be more appropriate. The ABCD schema already makes use of several enumerations in a similar context (e.g. RecordBasisEnum and SexCodeEnum). Consideration should be given to ensure consistency of values in both DwC and ABCD data sets.

The biodiversity informatics community is also concerned about describing and accessing a variety of ancillary information that might be associated with a particular specimen or occurrence record. This ancillary information will often be critical for analyzing processes

[32] http://www.bioontology.org/wiki/index.php/OboInOwl:Main_Page

[33] http://www.tdwg.org/activities/abcd/

[34] http://rs.tdwg.org/dwc/index.htm

[35] http://dublincore.org/

[36] http://rs.tdwg.org/dwc/terms/index.htm

[37] http://vocabularies.gbif.org/vocabularies/type_vocabulary

and patterns in biodiversity, and enlarges the scope of potentially relevant data to include a broad range of observed measurements about the biotic and abiotic aspects of the environment. For example, when examining patterns in the global abundance of some taxon, information regarding the co-situated precipitation, frost-days, soil type, land use, etc., could all be important parameters for analysis. Thus, the biodiversity community's informatics integration needs ultimately to converge with those of other earth and environmental sciences that rely on multi-disciplinary data for integrated or holistic understanding.

The **TDWG Vocabularies** are a set of OWL ontologies originally designed for use in LSID metadata resolution. Some 30 ontologies in the set can be roughly divided into six categories: Humans and Human Institutions, Taxonomy, Occurrence, Time and Space, Media, and Foundations. These and some further detail may be found in Appendix 6.1 TDWG Vocabularies.

The Vocabularies are hosted at a Google code site[38] as part of an effort to make them more accessible. That re-hosting is part of an expected redesign of the Vocabularies, for which, however, no precise plan has emerged. Elsewhere in this report we mention a few of them that are currently known to be in use. Some of the vocabularies have garnered greater traction than others. In particular, the vocabularies under the Taxonomy heading have been utilised in several applications, and the other vocabularies tend to be useful to the degree to which they are used by the Taxonomy vocabularies. None of the vocabularies has been approved by the TDWG standards committees. Although the vocabularies use OWL, their semantics are limited to the widespread use of property domain and range restrictions.

The precise relationship of the Global Names Architecture (see Section 2.2.3) to the TDWG vocabularies and their proposed but unscheduled redesign has not yet emerged.

The **GBIF Vocabularies**[39] are a set of community-supported vocabularies and Darwin Core extensions, with support for multilingual names and definitions and served through a ScratchPads user interface supporting REST XML web services and downloadable tab-delimited and CSV files. Available vocabularies include basisOfRecord (for DwC), Country, Language, Life Form, Life Stage, IUCN Habitats and others. At least some of these vocabularies are under active management. In the absence of recommendations on the use of such vocabularies within DwC and ABCD data sets, it is unclear how widely these are being used in actual data sets. Their broader adoption would serve to clarify the interoperability of different data sets and could provide a foundation for future semantic inference based on these data. Of particular importance is that this resource is implemented as a community mechanism for the management of vocabularies ("the GBIF Vocabularies Server" - an existing KOS facility that we discuss later both in general and as to this solution).

The **Ecological Metadata Language (EML)**[40] is a mature specification, with an XML-Schema implementation, addressing many aspects of the description of ecological data. Among these are the data field names and datatypes, research methodology for the observations, literature citation, and access control. The EML parent project also offers several tools

[38] http://code.google.com/p/tdwg-ontology/

[39] http://vocabularies.gbif.org/

[40] http://knb.ecoinformatics.org/software/eml/

supporting EML use: the **Morpho**[41] data management platform and the **Metacat**[42] data catalogue platform.

The **TDWG Structured Descriptive Data (SDD)**[43] standard is designed as an exchange standard primarily for morphological data sets and diagnostic keys. Using it for semantic applications is possible but cumbersome because SDD documents traditionally carry the entity and attribute definitions and their use in a single document. This can needlessly add to the difficulty of harmonizing vocabularies in different SDD documents, but it is not insurmountable.

SDD is in use in a number of standalone identification systems, as an exchange format in **KeyToNature**[44], and the **IdentifyLife**[45] project which is working to provide a framework for interoperable integration of diagnostic keys from many sources and formats, including SDD, Lucid,[46] Delta[47] and dichotomous keys. IdentifyLife has begun a "key to all life" project with partners including ALA, EOL and the Moore Foundation.

The **TDWG Species Profile Model (SPM)**[48] is presently mainly a set of widely used biology concepts. One application[49] describes a service that extracts taxonomic descriptions from biosystematics data and serves it encoded as RDF valid as SPM individuals. That is known to be harvested by the EOL for integration with species pages.

SPM provides about thirty categories given as subclasses of a class named *InfoItem* to express different concerns of biology, such as Cytology, MolecularBiology, Ecology, Behaviour, etc. None presently have properties beyond those of the InfoItem base class and none have even informal definitions that would allow the expression of a few common properties in plain text or controlled vocabularies. The base properties support the ability to describe the content of the InfoItem in plain text and provide for spatio-temporal or taxonomic contexts in which the content is valid.

SPM has been used in EOL and other contexts but there is a strong perception (e.g., in recent work as part of the Australian Taxonomy Research & Information Network (TRIN)[50] project and the above mentioned Plazi-EOL Project) that it is unduly constrictive and should not be conceived as a single vocabulary but as a pluggable model for indicating terms from any of many concept vocabularies (e.g., taxon-specific vocabularies such as the previously mentioned OBO format anatomy ontologies) that may classify a given InfoItem.

---

[41] http://knb.ecoinformatics.org/morphoportal.jsp

[42] http://knb.ecoinformatics.org/software/metacat/MetacatAdministratorGuide.pdf

[43] http://wiki.tdwg.org/twiki/bin/view/SDD/WebHome

[44] http://keytonature.eu

[45] http://identifylife.org

[46] http://www.lucidcentral.com/

[47] http://delta-intkey.com/

[48] http://wiki.tdwg.org/twiki/bin/view/SPM/WebHome

[49] http://wiki.tdwg.org/twiki/bin/view/SPM/PlaziEOLProject

[50] http://www.taxonomy.org.au/

The **Convention on Biological Diversity (CBD)**[51] Controlled Vocabulary is a somewhat dormant, but extensive, SKOS-like vocabulary designed for semantic searches on CBD documents.

The NBII Biocomplexity Thesaurus (BCT)[52] is a well-maintained SKOS thesaurus in use at the U.S. National Biological Information Infrastructure. It is the merger of five thesauri, one of which, notably, addresses sociological vocabulary. The BCT is accessible with a web interface and also via a WSDL-based web service. The web site also contains brief descriptions, with links, of approximately 150 other biodiversity and ecological terminologies.

The **World Wide Web Consortium (W3C)** maintains many specifications and interest groups related to the semantic web[53]. The most relevant of the W3C interest groups is the Semantic Web Health Care and Life Sciences (HCLS) Interest Group[54] . Although its charter covers biological science in general, most HCLS activity is dedicated to molecular biology and biomedicine. An important exception may be its Task Group on Scientific Discourse.

Habitat Classifications. Many regional, national, and sub-national governments and NGOs publish habitat classification schemes for conservation and decision support purposes and to promote data integration of habitat data. One typical one is the pan-European **EUNIS Habitat types classification**[55] a comprehensive pan-European habitat classification. There appears to be no KOS framework for describing habitat classification systems, so that comparing across them may be quite difficult.

The **Environment Ontology (EnvO)**[56] is a mature OBO format ontology that is supported under the umbrella of the OBO Foundry. It aims to support the semantically consistent description of, and computational reasoning over, observed environments associated with biological data of any organism or biological sample.

Literature: TaxPub[57] is an extension to the NLM/NCBI Journal Article Tag Suite, under development by Plazi[58] with cooperation from the U.S. National Center for Biotechnology Information. The TaxPub extension aims to enable the semantic enrichment of new taxonomic literature through XML markup. GBIF has previously managed a Plazi project for the service of SPM taxon descriptions extracted from legacy publications.

---

[51] http://www.cbd.int/doc/cbd-voc.aspx?id=5810

[52] http://www.nbii.gov/portal/server.pt/community/biocomplexity_thesaurus/578/about_biocomplexity_thesaurus/1658

[53] http://www.w3.org/standards/semanticweb/

[54] http://esw.w3.org/HCLSIG

[55] http://eunis.eea.europa.eu/habitats.jsp

[56] http://www.environmentontology.org/

[57] http://sourceforge.net/projects/taxpub/files/

[58] http://www.plazi.org

2.2.3 Data Providers

Taxonomic and Nomenclatural Catalogues. There are a number of standard catalogues of the names in use for organisms (both scientific and common names), including links to metadata on associated publications, and expert-curated views of the taxonomy and classification for different groups of species, are foundational resources for biodiversity informatics. Nomenclatural databases such as IPNI[59], Index Fungorum[60] and ZooBank[61] can serve to promote consistent representation and use of information on names and naming events. Key resources such as the Catalogue of Life[62] (including Species 2000[63] and ITIS[64]), WoRMS,[65] Species Fungorum[66] and many national and regional data sets provide taxonomic judgments on the number and names of accepted species and their organisation in a consensus classification.

Many of these resources have been exposed via web query interfaces and a number have been shared using LSIDs and the TDWG vocabularies for TaxonNames and TaxonConcepts. On a larger scale, many of the projects involved have been working together under the title of the **Global Names Architecture**[67] to develop a consistent model for sharing, discovering and using these data, including appropriate representation and mapping between alternative taxonomic opinions. This work is essential for the future development of biodiversity informatics. A robust suite of tools and services to discover any published name and its relationship to taxon concepts presented in key taxonomic resources would significantly enhance the interoperability of biodiversity data at all levels.

Gazetteers - a wide range of gazetteers are in use in support of biodiversity informatics. Some of these are listed by the **BioGeomancer**[68] project on its Gazetteers page[69]. The Open Geospatial Consortium[70] (OGC) has defined a Gazetteer Service profile of the OGC Web Feature Service specification. Presentation of gazetteers using this profile can allow disparate national or international resources to be used interchangeably within biodiversity applications. Many (particularly botanical) projects have made use of the TDWG World Geographical Scheme for Recording Plant Distributions[71] as a hierarchical gazetteer of regions for recording species distributions. The **Gaz Project**[72] represents a first step towards an open source gazetteer, constructed on ontological principles, that describes places and place names and the relations between them.

---

[59] http://www.ipni.org/

[60] http://www.indexfungorum.org/

[61] http://www.zoobank.org/

[62] http://www.catalogueoflife.org/

[63] http://www.sp2000.org/

[64] http://www.itis.gov/

[65] http://www.marinespecies.org/

[66] http://www.speciesfungorum.org/

[67] http://www.globalnames.org/

[68] http://www.biogeomancer.org/

[69] http://www.biogeomancer.org/bg_library/links/gazeteers/

[70] http://www.opengeospatial.org/

[71] http://www.nhm.ac.uk/hosted_sites/tdwg/geo2.htm

[72] http://gensc.org/gc_wiki/index.php/GAZ_Project

Gazetteers differ in significant ways from some of the other KOS components listed here, in that their use in biodiversity informatics relates particularly to their role in interpreting textual locality information associated with specimens, and in mapping this text to coordinate- or polygon-based geospatial representations. For most applications, these numerical representations are more generally useful than the original locality name, although the latter is obviously an essential component of the metadata for the specimen and the justification for the interpreted values. In most cases, subsequent discovery of data records associated with a gazetteer location will be mediated through mapping the gazetteer terms to coordinates and searching for records occurring within a suitable geospatial envelope. It is not to be expected that species occurrence records will be annotated with a comprehensive set of gazetteer terms to represent all of the possible names for the associated locality.

Gazetteers are therefore highly valuable both in digitisation and in assisting users to discover relevant data, but are less likely to be used consistently in data interchange. Within individual databases, records are likely to be associated with a range of gazetteer terms of particular interest to the users of that database.

As a consequence, it makes sense for the biodiversity informatics community to focus on appropriate standardised representations of localities defined using coordinates, datums, precision and uncertainty. The IETF scheme for URIs for geographic locations[73] may provide a suitable basis for this. Tools and services would be required to build resolvable identifiers based on this model and hence to ensure that they could be used consistently in RDF and LOD applications.

The Biodiversity Collections Index (BCI)[74] is a collaboration established by GBIF, TDWG and the Royal Botanical Garden Edinburgh to integrate metadata on natural history collections from the various catalogues which hold such information for different communities, such as **Index Herbariorum (IH)**[75], **Insect and Spider Collections of the World (ISCW)**[76] and the **BioCASE metadata database**[77]. These catalogues were in many cases developed before the rise of web technologies and present short text strings as identifier codes for the various collections (these codes being specific to a given catalogue). BCI presents the contents of these resources in a consistent fashion and offers LSIDs and HTTP URIs for accessing RDF metadata for each collection. These identifiers are in use by various projects, including the Atlas of Living Australia and GBIF, as a means to standardise references to collections. The key challenge to maintaining BCI is the difficulty of maintaining currency and comprehensiveness for the collection metadata. The **Barcode of Life Initiative**[78] has also developed a repository to store mappings between historical institution codes and collection codes and associated web-queryable databases for retrieving specimen records. This repository is intended to hold less metadata than BCI and is in many ways complementary to it, allowing the Barcode of Life Initiative to maintain associations between GenBank sequences and the originating voucher material.

[73] http://tools.ietf.org/html/rfc5870

[74] http://www.biodiversitycollectionsindex.org/static/index.html

[75] http://sciweb.nybg.org/science2/IndexHerbariorum.asp

[76] http://hbs.bishopmuseum.org/codens/codens-r-us.html

[77] http://www.biocase.org/whats_biocase/meta_net.shtml

[78] http://www.biorepositories.org/

The **International Long Term Ecological Research (ILTER)**[79] facility is a network of networks worldwide that gather and serve ecological data and support mechanisms for its integration; notably the ILTER operates an integrated Metacat instance that is intended to link the Metacats of its members.

The **Global Invasive Species Information Network** (GISIN)[80] has developed the GISIN Protocol[81] for exchange of information on invasive species. This includes standard terms and, in some cases, enumerated values for relevant concepts.

The National Center for Biomedical Ontology (NCBO) **BioPortal**[82] provides an open repository of biomedical ontologies, including those of the OBO Foundry. It enables browsing, visualisation, search across ontologies, mappings, structured comments, and term requests. The BioPortal platform itself is not specific to biomedicine and is available for installation for other domains. The platform is the subject of an adoption recommendation later in this report, described with more detail.

The **UN Food and Agriculture Organization (FAO)** operates the Agricultural Information Management Standards[83] web site which points to a number of resources, including the AGROVOC thesaurus[84] a "multilingual, structured and controlled vocabulary designed to cover the terminology of all subject fields in agriculture, forestry, fisheries, food and related domains (e.g. environment)" and a registry of KOSs relevant to FAO concerns. **Bioversity International** has developed "Descriptor lists" and derived standards to describe, store, and manage information about plant resources to enable international information sharing. Descriptors form a vocabulary specialised to each crop plant to enable information sharing for crops covered under ANNEX 1 of the International Treaty of Plant Genetic Resources for Food and Agriculture Organization[85]. The various documentation standards share a common core in the form of FAO/IPGRI Multi-crop passport descriptors[86] compatible with the FAO World Information and Early Warning System (WIEWS) on plant genetic resources. These are the basis of major information platforms such as GENESYS[87] and EURISCO[88]. The vocabularies are divided into categories, some of which are common across species. Descriptor concepts include terms needed for the specimen management, for description of the collection event and environment, for the management of the crop, for morphological description, etc. Terminology included is the result of global consultations amongst crop specific experts. Bioversity International is a partner in the **Consultative Group on International Agricultural Research (CGIAR)**

[79] http://www.ilternet.edu/

[80] http://www.gisinetwork.org/

[81] http://www.niiss.org/cwis438/websites/GISINDirectory/tech/Protocol_Home.php?WebSiteID=4

[82] http://bioportal.bioontology.org/

[83] http://aims.fao.org/home

[84] http://aims.fao.org/agrovoc-thesaurus-ontology

[85] http://www.bioversityinternational.org/policy_law/international_seed_treaty.html

[86] http://www.bioversityinternational.org/nc/publications/publication/issue/multicrop_passport_descriptors.html

[87] http://www.genesys-pgr.org/

[88] http://eurisco.ecpgr.org/static/index.html

whose Information and Communication Technology - Knowledge Management (ICT-KM) [89] Program has developed the Knowledge Sharing Toolkit (KSTK)[90]. The toolkit does not have any obvious support for any primary biodiversity data web services operated by the CGIAR partners or FAO. The principals in the development of KSTK have deployed the semantically enhanced **Agropedia**[91] apparently based on modelling with Concept Maps.

2.2.4 Projects, platforms, and practices

The **Open Biological and Biomedical Ontologies (OBO) Foundry**[92] is a well supported effort whose goal is to create a suite of community-developed orthogonal interoperable reference ontologies in the biomedical domain. The OBO Foundry standards and procedures, as well as several upper ontologies that domain ontologies can draw upon, have led to a number of OBO format ontologies useful for application wider than biomedical informatics, most notably for phylogenetic studies. Central among these is the **Phenotypic Quality Ontology (PATO)**[93]**,** which provides a framework for a number of OBO format anatomy ontologies, including those for Hymenoptera, Diptera, and Humans. PATO could also be used for species descriptive data, including in terminology sections of descriptions encoded in the TDWG Structured Descriptive Data (SDD)[94] schema, though we are unaware of such use.

The **Phenoscape**[95] project uses the EQ approach to expose characters and character states from the systematics literature to large-scale computational analysis. This includes integrating natural phenotypes with mutant phenotypes of model organisms, with the aim to generate hypotheses about the genetic causes of evolutionary character transitions. Phenoscape uses some Darwin Core terms to identify the specimen supporting a phenotype observation. Beyond the shared use of biodiversity vocabularies, the Phenoscape work also shows the potential of applying ontologies and formal knowledge representation techniques to descriptive biological data in general.

The **Hymenopteran Anatomy Ontology (HAO)**[96] project uses EQ ontologies to standardise descriptions of phenotype diversity among Hymenopteran insects across the taxonomic and systematics literature. The **Phenex**[97] application is designed to annotate character matrices with ontology terms. It builds on Phenote[98] and OBOEdit. The **Phenote** project provides software for phenotype annotation using the EQ model.

[89] http://ictkm.cgiar.org/index.php

[90] http://www.kstoolkit.org/

[91] http://agropedia.iitk.ac.in/

[92] http://www.obofoundry.org/

[93] http://obofoundry.org/wiki/index.php/PATO:Main_Page

[94] http://wiki.tdwg.org/SDD

[95] http://phenoscape.org

[96] http://hymao.org

[97] https://www.phenoscape.org/wiki/Phenex

[98] http://www.phenote.org/about.shtml

The recently funded **Phenotype Ontologies Research Coordination Network**[99] is an organisation of investigators and projects interested in the representation of morphology, behaviour, and other phenotypic traits through the use of ontologies.

Linked Data[100] is a set of practices in the Semantic Web community leveraging the use of HTTP URIs as identifiers for all things (whether "information" or "non-information resources") in order to facilitate integration of data across web accessible resources. Originally laid out in Berners-Lee's note *DesignIssues/LinkedData,*[101] a number of tools now exist to exploit these practices for semantically based-discovery. As examples, see the **OpenLink Data Explorer**[102] and offer it "Quercus alba" or the SPARQL query[103] in the **GeoSpecies** project[104] based on a small purpose-built ontology[105] of mosquito-borne human pathogens. Particular attention has been focused on the Linking Open Data (LOD)**[106]** cloud , a set of well-known datasets exposed by the community as part of the W3C-housed Linking Open Data project[107].

There is scattered biodiversity interest in Linked Data and the LOD cloud, but very little science-based semantics is on the table to support semantic links other than those based on taxonomy and geolocation. Recent discussions[108] on the TDWG-Content mailing list occasionally dip into LOD, either in advocacy or examples. It is likely that anatomy ontologies are easily exploited for discovery with Linked Data tools, but formal anatomy ontologies seem to be limited to a few specific groups such as those mentioned above in OBO format and the Foundational Model of [Human] Anatomy (FMA), which has several OWL representations. **Bio2RDF**[109] converts several dozen genomics databases to about 30 billion triples, some of which are available through the LOD cloud and all are available through the OpenLinkSoftware linked data mashup[110]. The **l**inkage is based on five ontologies[111] focused mainly on molecular biology. Reuse in one's data services of the URIs of other publishers' URIs for their data allows Linked Data applications to automatically find these other resources. Further, when even minimal scientifically useful ontologies are describing the data with widely reused URIs, then discovery, integration, and retrieval can exploit those ontologies throughout the cloud of links.

The current LOD cloud is small compared to biodiversity data on some measures. For example, at this writing, LOD statistics[112] reveal only 42 bioscience datasets (including some of the Bio2RDF views on important molecular biology data) holding 2.7 billion

[99] http://phenotypercn.org/

[100] http://linkeddata.org/

[101] http://www.w3.org/DesignIssues/LinkedData.html

[102] http://linkeddata.uriburner.com/ode/

[103] http://about.geospecies.org/sparql.xhtml#example_8

[104] http://about.geospecies.org/index.htm

[105] http://rdf.geospecies.org/ont/families/wQViY/wQViY_ontology.owl#

[106] http://www.ckan.net/group/lodcloud

[107] http://esw.w3.org/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[108] http://lists.tdwg.org/pipermail/tdwg-content/2010-November/subject.html

[109] http://bio2rdf.blogspot.com/

[110] http://lod.openlinksw.com/

[111] http://sourceforge.net/apps/mediawiki/bio2rdf/index.php?title=Ontologies

[112] http://www4.wiwiss.fu-berlin.de/lodcloud/state/#domains

triples, both of which are an order or two of magnitude smaller than would be required for the data behind the GBIF occurrence cache alone.

The breadth of the LOD cloud presents both opportunities and challenges for GBIF. For example, the cloud presently holds 25 governmental datasets with 11 billion triples. To the extent that governments and conservation NGOs embrace it, semantically enabled links between primary biodiversity data and social impacts may become very useful for public policy decision support use of biodiversity data.

NeON[113] is an EU project for the management and application of ontologies. It ended in March 2010 but seems to be releasing a final (?) version of the NeONToolkit (NTK)[114] for ontology life-cycle management. NTK has about 45 plugins and its website mentions 12 projects based on, or originated in, NeON. The only biological application seems to be the FAO Network of Fisheries Ontologies[115], which remains in draft form. One of the corporate NeON partners, Ontoprise, provides a commercial version of the Semantic MediaWiki (SMW) extensions to MediaWiki and there may be some support for SMW in the NTK. Most of the NeON community mechanisms seem somewhat dormant at this writing.

The **NCBO BioPortal** software is both a platform for ontology repositories and an instance of one. We discuss it extensively elsewhere since it is the subject of a recommendation for adoption.

The Open **Ontology Repository (OOR)**[116] Initiative is a two-year old community project whose charter "is to promote the global use and sharing of ontologies by:

1. establishing a hosted registry-repository;
2. enabling and facilitating open, federated, collaborative ontology repositories, and
3. establishing best practices for expressing interoperable ontology and taxonomy work in registry-repositories." [117]

The OOR high level requirements[118] set forth technical and community requirements identified to meet the charter goals. A test implementation[119] based on the NCBO BioPortal platform has been deployed. The Spatial Ontology Community of Practice (SOCoP) is a newly funded project using OOR.

Semtools[120], a project funded under the U.S. National Science Foundation Advances in Biological Informatics program (NSF ABI), is testing how observational data models can be used to semantically enhance understanding of ecological (including biodiversity) data. It builds upon the metadata editor, Morpho[121] that is used to create, store, and query

[113] http://www.neon-project.org

[114] http://neon-toolkit.org/wiki/Main_Page

[115] http://aims.fao.org/en/website/Fisheries-ontologies-/sub2

[116] http://ontolog.cim3.net/cgi-bin/wiki.pl?OpenOntologyRepository

[117] http://ontolog.cim3.net/cgi-bin/wiki.pl?OpenOntologyRepository#nid17ZH

[118] http://ontolog.cim3.net/cgi-bin/wiki.pl?OpenOntologyRepository_Requirement

[119] http://oor-01f.cim3.net/

[120] https://semtools.ecoinformatics.org/

[121] http://knb.ecoinformatics.org/morphoportal.jsp

Ecological Metadata Language (EML) metadata. Semtools is implementing interfaces onto EML-described datasets that allow semantic annotation of the referenced raw data by using an observational data model, the Extensible Observation Ontology (OBOE)[122]. The OBOE OWL ontology structures annotations of the metadata in ways that enable linking domain ontology terms to the data via a well-specified observational structure. By leveraging standard inference engines, such as Pellet[123], the observational data model and associated OWL ontologies enable a number of unique reasoning services that should prove useful for investigators. These include finding data sets where specified measurements co-occur, facilitating powerful search through term expansion, verifying whether queries are ontologically consistent, and clarifying when multiple observations are taken from one particular specimen or instance.

IdentifyLife[124] is a platform to support sharing of metadata relating to identification keys and to facilitate reuse of descriptive character definitions and of character data for individual taxa. IdentifyLife will expose web services for discovery and reuse of these data.

Atlas of Living Australia (ALA)[125] is a national infrastructure project to integrate available information sources relating to Australian biodiversity. The project has been following a Linked Open Data approach, based around key KOS resource for nomenclature and taxonomy (Australian Plant Name Index, Australian Plant Census, Australian Faunal Directory, Catalogue of Life, Interim Register of Marine and Non-marine Genera – (IRMNG), geography (national gazetteers, World Database of Protected Areas, Interim Biogeographic Regionalisation of Australia, Interim Marine and Coastal Regionalisation of Australia), national threatened species lists, etc. Other classes of information are being handled using LOD to maximise interlinkage and connections with these core references. The ALA is in particular working with the IdentifyLife project to link data from identification tools (characters and states) into this framework. Where available and appropriate, the ALA is using TDWG vocabularies (particularly Darwin Core, ABCD, the TaxonName and TaxonConcept ontologies and SDD) to maximise interoperability with other projects. The focus at this stage is on improved data integration and data cleansing (e.g., through use of the marine/non-marine property for taxa from IRMNG) and on improved search and discovery. Further semantic exploration of possible use for integrated linked open data exposed through ALA web services is being deferred.

VIBRANT[126]**,** the *Virtual Biodiversity Research and Access Network for Taxonomy*, is a major EU project just begun to build community mechanisms and cyberinfrastructure for virtual communities of biodiversity researchers. By its design and specification, its cyberinfrastructure will be built on ScratchPads[127], which presently defines itself as "a social networking application that enables communities of researchers to manage, share and publish taxonomic data online." GBIF is a VIBRANT partner.

[122] Madin et al. 2007, http://dx.doi.org/10.1016/j.ecoinf.2007.05.004

[123] http://clarkparsia.com/pellet/

[124] http://www.identifylife.org/

[125] http://www.ala.org.au/

[126] http://vbrant.eu/content/vibrant-supporting-biodiversity-research-communities

[127] http://scratchpads.eu/

Annotation: A TDWG Annotation Interest Group has been formed with core members from the Filtered Push[128] project, GBIF, and ALA.

Morphbank[129] is a mature repository for depositing and retrieving specimen and other biodiversity images. It is now collaborating with the **Morphster**[130] project to allow images to serve as annotations of anatomical ontologies and conversely to allow ontology-driven searches for anatomical images.

Multimedia: A biodiversity multimedia metadata vocabulary is nearing the end of the TDWG review process. It is the product of a joint TDWG/GBIF Multimedia Resources Metadata Task Group (MRTG)[131]. When accepted, it will be known as the Audubon Core[132].

LifeWatch[133] is a major EU initiative in its planning stages to address many aspects of biodiversity data, its mobilisation and use. It will no doubt require all the aspects of KOS mentioned in this report. GBIF's existing Memorandum of Cooperation with LifeWatch positions it to coordinate its KOS efforts with any of those of LifeWatch.

GEMET[134] is a multilingual (some 29 languages) environmental thesaurus used as an indexing, retrieval and controlled vocabulary tool by the European Environment Agency.

The **NASA Global Change Master Directory (GCMD)** Science Keywords[135] is a set of hierarchically arranged keyword vocabularies covering the earth sciences and including biodiversity related terms.

Semantically enhanced wikis have been proposed or deployed for biological science KOS[136] and there is discussion of their possible use in the VIBRANT project[137].

2.2.4.1 Related projects in other disciplines

The **Marine Medatata Interoperability Initiative (MMI)**[138] produces metadata standards and tools for marine science data. Particularly of interest for biodiversity data are its efforts developing observation ontologies and sensor ontologies. In addition, MMI tracks related efforts on its community news pages and maintains extensive lists of KOS resources

---

[128] http://etaxonomy.org/mw/FilteredPush

[129] http://www.morphbank.net/

[130] http://morphster.org/

[131] http://www2.gbif.org/MRTG-Recommendations-29-09-2008.pdf

[132] http://www.keytonature.eu/wiki/MRTG_v1.0

[133] http://www.lifewatch.eu/

[134] http://www.eionet.europa.eu/gemet

[135] http://gcmd.nasa.gov/Resources/valids/archives/keyword_list.html

[136] e.g. http://bowiki.net/wiki, http://code.google.com/p/sbpipeline/wiki/SbWiki

[137] Comments of G. Hagedorn in response to an early draft of this paper; debate of R. Page and V. Smith in http://iphylo.blogspot.com/2009/01/wikis-versus-scratchpads.html and http://vsmith.info/Breaking-Barriers

[138] http://marinemetadata.org/

of interest to its participants (e.g., Conventions[139], Tools[140], Initiatives[141], Ontologies and Thesauri[142], lists, and others).

CLARIN, ISOcat, ISO 12620. Addressing data-centric concerns of linguistics, including those of ISO 12620[143] (computer applications in terminology), the Common Language Resources and Technology Infrastructure (CLARIN)[144] project is a large-scale pan-European collaborative effort to create e-humanities tools for linguistics. The most advanced effort of CLARIN is an implementation of ISO 12620 called ISOcat.[145] In its domain (Linguistics), ISOcat shares some functionality about vocabulary term management with the GBIF Vocabularies Server, albeit perhaps with more extensive structure such as is exhibited in the BioPortal platform. Although ISOcat's community mechanisms appear to be relatively new (e.g., its forum[146] has only 13 posts, 12 of them support issues, and the software is in beta), they are very well structured as to term management. CLARIN itself is nearing the end of a preparatory phase with design to begin in 2011 and full deployment in 2016. If its utility proves high to its community, ISOcat is likely to become independent of CLARIN's evolution. An early document[147] and personal email to the convener of this report claim that the software is not tied to the vocabulary of any specific discipline. Further study is needed to ascertain whether it would complement or overlap BioPortal and the GBIF Vocabularies Server.

The **Science Commons Term Broker**[148] is an implementation for neurosciences of an open source, configurable, ontology term broker platform of neurocommons.org. The implementation is coupled to a document annotation system and allows the annotator to find suitable terms for their annotation, or request the ontology managers to create a new one. It is only recently released and has few, if any, users but may be worthy of study.

2.2.5 Life cycle tools

Integrated KOS tools covering the entire lifecycle—design through to management—are the subject of advanced informatics development projects, some of which GBIF is a partner in. The most mature and widely used tools for complex KOS resources are ontology editors and browsers, and we here mention a few of the better known ones.

Protégé[149] is the most mature and widely adopted ontology editor in use, with over 160,000 registered users and nearly 200 self-described projects[150] using it. It is an open

[139] http://marinemetadata.org/conventions

[140] http://marinemetadata.org/tools

[141] http://marinemetadata.org/intitiatives

[142] http://marinemetadata.org/conventions/ontologies-thesauri

[143] http://www.iso.org/iso/catalogue_detail.htm?csnumber=2517

[144] http://www.clarin.eu

[145] http://www.isocat.org/

[146] http://www.isocat.org/forum/index.php

[147] http://www.clarin.eu/files/concept_registry-CLARIN-ShortGuide.pdf

[148] http://neurocommons.org/page/Ontological_term_broker

[149] http://Protégéwiki.stanford.edu/wiki/Main_Page

[150] http://Protégé.cim3.net/cgi-bin/wiki.pl?ProjectsThatUseProtégé

source platform with a documented plugin API that has led its developer community to provide wide functionality with over 80 plugins already contributed[151], including graphical tools, OBO format editing support, reasoners and others. Its current version now in late beta handles OWL2. Like most ontology editors, Protégé anecdotally has a reputation among many (non computer) scientists as being difficult to use, but a web-based version now emerging is intended to address that. CollaborativeProtégé[152] is a Protégé extension that supports collaborative editing and annotation of ontology components and ontology changes.

TopBraid Composer(TBC)[153] is an OWL editor component of a commercial suite of modelling tools. Its free edition has SPARQL support and a friendly forms-based ontology editor, but has limited functionality compared to the entire suite. TBC is an Eclipse plugin so can be run standalone or within Eclipse. This makes it possible to use other Eclipse plugin support for some life cycle aspects such as version control using popular source code repository systems.

OBO-Edit[154] is an editor optimised for OBO format ontologies, one of whose aims is ease of use. It is funded by the Gene Ontology Consortium, so even with a small developer community it is likely to be well maintained as long as it is deemed useful.

CmapTools[155] is a free software suite for collaborative construction, sharing and publishing of knowledge models represented as concept maps. Its web site claims it is used worldwide by millions of users, and is available in over 15 languages. Concept maps are a gentle way to introduce general principles of knowledge representation, and the software is very easy to use, at least for individuals (some of the emphasis is on collaboration, which we have not evaluated). There is a full-blown OWL editor[156] on top of the suite, but it is a single-developer project and possibly with small user community. Because the base is a completely graphically-based system, the OWL editor is best suited for small ontologies.

The **Manchester University Ontology Browser**[157] is a web-based OWL ontology browser with a RESTful interface that should make it easy to invoke from other tools. For example, the MMI BioPortal implementation of its Ontology Registry and Repository does such invocation.

2.3 Gaps in Current Biodiversity KOS

The survey revealed no gaps as identified by the community. In retrospect, it probably could not have, because in pursuit of brevity, the questions invited respondents to say what do they do now or hope to do in the future and had no way to identify which was being answered. Below follows a synthesis of the authors' views of the gaps.

---

[151] http://Protégéwiki.stanford.edu/wiki/Protégé_Plugin_Library

[152] http://Protégéwiki.stanford.edu/wiki/Collaborative_Protégé

[153] http://www.topquadrant.com/products/TB_Composer.html

[154] http://oboedit.org/

[155] http://cmap.ihmc.us/

[156] http://www.ihmc.us/groups/coe/

[157] http://owl.cs.manchester.ac.uk/browser/

2.3.1.1 General gaps

In general, resource discovery is haphazard. Google is inadequate. For example, a Google search for "Ant bibliography" produces as the third item a reference to the extensive FORMIS: Master Bibliography of Ant Literature[158], but a search for "Hymenoptera bibliography" or even "Hymenoptera Systematics Bibliography" did not find FORMIS until the fifth page of Google returns, and only found two references to it in the first 10 pages. A Bing search had similar results.

Existing ontology life-cycle tools are mostly too difficult for domain scientists to use, and often are not well integrated.

The grey biodiversity literature (e.g., government documents and conservation NGO documents) appears to have no uniform discovery mechanisms, but is likely very valuable for conservation and education applications. Individual governmental agencies (e.g., the NBII Metadata clearinghouse[159]) may operate broad literature of sister agencies publications but we are unaware of any international "catalogue of catalogues".

There is no systematic community approach to setting and documenting KOS goals, e.g., by providing use case and competency question libraries.

2.3.1.2 Gap in breadth

There is no harmonised model for scientific observations. Contributions to filling this gap represents one of the biggest opportunities for GBIF to leverage its own expertise along with those of other efforts (e.g., those of the Marine Metadata Interoperability Initiative, MMI) to build more widely useful earth sciences (broadly understood) information systems necessary not only for science, but also for environmental decision support and educational uses.

The notion of a generalised model for scientific observations has evolved independently within a number of earth science disciplines, largely over the past half decade, as each discipline was struggling with how to achieve data interoperability within its own domain. These communities included the biodiversity sciences (from a fieldwork, occurrence-based, tradition), ecology, evolution, geospatial sciences, hydrology, oceanography, solar-terrestrial physics, and socioecology, among others. In each case, domain-oriented informatics experts had perceived utility from adopting a core model for scientific observations and measurements that could be flexibly linked to domain-sanctioned controlled vocabularies—e.g., terms drawn from OWL ontologies—in order to achieve highly flexible and potentially semantically powerful data interoperability. The basic structure of a scientific observation involves clarifying that "measurements" are the documented values of "properties" (which can include "classifying" or "naming" as well as counting or assigning metric values to quantities) that are characteristics of some "things" or "entities" (or processes). This formalisation requires an ontological commitment, since philosophers are still debating the fundamental status of things as such, as well as the epistemological basis for asserting objective observations of properties of distinct things, using these to differentiate instances, and placing these instances into categories that reflect sets of "natural types". Nevertheless, there are strong practical advantages to moving forward with a harmonised model for scientific observations.

By providing a conceptualisation of a datum as the measurement of some characteristic of something, observational data models encourage knowledge modellers to develop supportive domain ontologies that adequately and consistently describe the fundamental facets of an observation. Thus, these models ground ontologies at a highly relevant and

[158] http://www.ars.usda.gov/Research/docs.htm?docid=10003

[159] http://metadata.nbii.gov/clearinghouse/

fundamental level for scientific investigation—the atomic level found in scientific data sets. As communities grapple with how to construct KOS, including ontology development, the observational abstraction provides underlying scaffolding for describing specific observations, and enables these observations to be inter-related. That is, observational data models enable scientists to view a diversity of data structures as sets of associated measurements. This ability to explicitly and semantically contextualise how one observation is related to another (e.g. observations will almost always have a relevant context of time and place) is a valuable feature that is found in many observational models. But in many data sets, associated measurements are simply "joined" together in a row or tuple structure, and the nature of the inter-relationships is unspecified. Examples include nesting (e.g., plot within site) and other forms of containment (e.g., isopod on fish), and general context (e.g., sample from stream-water).

There are currently several activities to harmonise observational data models, and optimise possibilities for strong compatibility among these. The OBOE ontology and SONet (Scientific Observations Network)[160] are two examples of efforts to develop and explore how observations can be modelled, and there are activities within the TDWG Observation and Specimen Records Interest Group (OSR)[161] and its Observations Task Group[162], as well as some joint working groups involving NSF's DataNet projects—all of which are collaborating to develop shared observational data models, and exemplar use cases that emphasise data interoperability. Biodiversity use cases are well-represented in these efforts. The Open Geospatial Consortium (OGC Observations and Measurements (O&M)[163] model is particularly of interest in some sensor data communities (e.g., the Marine Metadata Interoperability Initiative - see Section 2.2.4.1), which has some connection to biodiversity observation projects at least through one active participant in SONet. In Europe, the SERONTO observation ontology[164] (which is informing the LifeWatch architecture)[165] shares many concerns with OBOE and the OGC O&M model. A recent workshop[166] at TDWG 2010 began to explore how these and other observation models could exchange data with one another.

It is not currently clear how these observational data models will interact with extensions of the Darwin Core standard. The biodiversity informatics community is embracing the challenge of documenting broader types of observational data that add context to records of taxon occurrences, but potential extensions to Darwin Core will certainly require incorporation of ontologically-defined terms and concepts that are better provided by experts in those distinct disciplines, whether phylogeneticists, soil scientists, climatologists, marine biologists, or other specialists. There are significant connection points between observational data models and current terms found in Darwin Core (as of 2009-12-07), most promisingly in the terms prefixed "measurement...", "associated...", and "dynamicProperties".

The Darwin Core approach, however, is primarily a growing set of identified attributes that can be documented for a given occurrence record. It does not have an overarching theoretical framework such as an observational data model, though it might readily adapt

[160] https://sonet.ecoinformatics.org/

[161] http://www.tdwg.org/activities/osr/

[162] http://wiki.tdwg.org/Observational

[163] http://www.opengeospatial.org/standards/om

[164] http://www.alter-net.info/SITE/UPLOAD/DOCUMENT/outputs/WPI6_2009_10_SERONTOCore.pdf

[165] http://www.slideshare.net/nichbuick/semantic-data-integration-of-biodiversity-data-with-the-seronto-ontology

[166] https://sonet.ecoinformatics.org/workshops/tdwg-2010-meeting

to one. Moreover, Darwin Core does not prescribe any clear distinction between the property being measured and the entity with which it is associated. It would be beneficial to all communities investigating biodiversity phenomenon if standards like Darwin Core were structured and extended in ways that integrate well with broader data interoperability efforts emerging in the earth and life sciences, of which observational data models are one prominent and highly relevant activity.

2.3.1.3 Specific gaps

Several specific gaps are recognisable in the resources we surveyed above:

- Darwin Core lacks an RDF representation supporting reasoning. Perhaps more importantly, despite its success as an exchange format for taxon occurrences, neither it nor ABCD have yet been fit in any kind of broader ontology of biodiversity knowledge
- Understanding, production, and exploitation of OBO format ontologies rest principally with the phyloinformatics and molecular branches of the biodiversity community, though such ontologies could be more widely exploited for species descriptions and for identification tools.
- Tools for extracting knowledge about species from biosystematics literature are at various states of maturity, and with weak sustainability models. The corresponding gap—coarseness of constraint schemas and lack of tools for semantic markup of born-digital taxonomic treatments—still imposes needs for Natural Language Processing (NLP) to provide extraction of species descriptive data. The growing corpus of born-digital scientific literature presents opportunities for development of tools and procedures that assist in semantic annotation early in their production, requiring no NLP on a corpus whose growth will dwarf existing publication.
- There is no current standard for representing species occurrence data with rich, machine-interpretable semantics, and that readily integrates with semantic standards for biological observation data related to species occurrences, such as ecological, molecular, trait, and phenotype observations. Specifically, there is no defined mapping or integration between Darwin Core records on one hand, and the OBOE, EQ, SERONTO or OGC O&M models on the other hand.
- Organised descriptive data in OBO format seems limited to a small collection of OBO anatomy ontologies.

3 Recommendations

3.1 GBIF participation in KOS standards development

Recommendation 1. Initiate and lead a joint TDWG/GBIF Task Group with at least the tasks below. Some of these tasks are of sufficient complexity that they may prove to require a separate Task Group be established. We have not attempted to prioritise these tasks, except that from our own discussions and those on many venues on the net, we know that the first item would address one of the biggest impediments to data discovery, exchange, and semantic enhancement.

a. Specify and implement a robust service for persistent unique identifier issuance.

b. Adopt or develop easy to use facilities for the creation of libraries of use cases and competency questions[167]. Such libraries should be accessible to, and interoperable with, other tools that GBIF deploys. For example, if such a library were based on standardised forms in a content management system such as Drupal, it is likely that it could easily exchange content with the structured Notes facility of BioPortal[168].

c. Make an OWL representation of Darwin Core. (Note that there is momentum for this already emerging in the TDWG-Content mailing list traffic.) Since DwC is designed for occurrence records, the future will likely prove it is only a small part of a larger synthesis of the kinds of ontologies we surveyed in Section 2.2.2. Such a synthesis will necessarily evolve as rapidly as biology itself, so it must be extensible and flexible. It should start with the establishment of systematic goals for biodiversity Knowledge Representation, and begin to lay out what is needed for a broad ontology of biodiversity, perhaps following Levin (2000)

d. Systematise goals for biodiversity Knowledge Representation. By example, initially select some target applications/use cases for KOS (e.g., the SPM categories). Assist the GNA design participants to formalise the requirements for the use, if any, of any of the current TDWG Ontologies, so that GNA and any TDWG Ontology redesign can co-evolve as needed.

e. Enhance SPM to develop a more flexible and inclusive model to serve the same goals as SPM. Shepherd the result through the TDWG adoption process.

f. Make an OWL representation of TDWG-SDD. As a first step, develop best practices to separate SDD Character and Taxon declarations from SDD Descriptions. Because descriptive data and their semantics are central for much of the knowledge about biodiversity, GBIF should where possible, facilitate work of domain experts to create and continually improve the prerequisite community ontologies, such as for species traits, habitats, life cycle, etc. There is opportunity to link a number of the collaborative projects listed in Section 2.2.4 that have a focus on descriptive data, including Atlas of Living Australia, IdentifyLife, KeyToNature, and the Phenotype Ontologies Research Coordination Network.

g. Make an OWL ontology for Taxonomic Treatments supporting knowledge extraction from both born-digital and retrospective systematic publishing. Identify and include adherents of e-nomenclature among the various Nomenclatural Commissions to help harmonise these efforts with the codes and their work practices.

---

[167] A brief overview of competency questions as an ontology design tool is found at http://marinemetadata.org/references/competencyquestionsoverview .

[168] Here we imagine tool interactions analogous to the way in which software project incident trackers can be automatically updated from activity in source code repositories.

h. Form a subgroup to identify the informatics and KOS requirements for social impact of biodiversity.

i. Form a subgroup on Linked Biodiversity Data with goal to foster semantics on links that are relevant to biodiversity. Examine and act on opportunities in GBIF data services to implement the practices of the Linked Data community.

j. Promote the widespread adoption of URI-based standard values for key Darwin Core attribute values (through the GBIF vocabularies and related activities) and work with the TDWG Darwin Core Task Group to develop appropriate mechanisms and documentation for their seamless use in Darwin Core data sets.

k. Evaluate the IETF scheme for URIs for Geographic Locations and provide recommendations for its use.

Recommendation 2. GBIF should represent biodiversity interests and perspectives in international observational data projects, several of which are broadly collaborative and actively seeking broad community involvement. In addition, GBIF should be receptive to helping implement technology solutions based on these standards, as they become available. In particular, GBIF should continue and emphasise its active participation in the Group On Earth Observations Biodiversity Observation Network (GEO BON), with active advocacy in GEO BON of KOS solutions to some of the barriers to progress GEO BON has identified.

Recommendation 3. GBIF should explore ways KOS can enhance the utility of providers of agricultural data. As an intergovernmental organisation GBIF is well positioned to, and should actively, explore KOS collaborative opportunities with non-GBIF agricultural data providers in the international community, such as FAO, CGIAR, Bioversity International and others.

3.1.1 Tool development and adoption

Recommendation 4. GBIF should deploy an instance of the BioPortal platform for biodiversity ontologies as a complement to the GBIF Vocabularies Server.

3.1.1.1 Rationale

The BioPortal platform has important features that make it useful for a Biodiversity Knowledge Organisation resource repository and directory as part of the life cycle management of ontologies. These include:

Functionality

- It supports ontology browsing, mapping between terms in ontologies in the repository, and notes that support commentary, new term proposal, and proposals for changes to terms.
- An ontology view materialisation feature in the current version provides for the publication and maintenance of ontologies derived from others (presently, by external applications), such as foundation ontologies, both by manually trimming and by use of an extension of SPARQL (Shaw, 2008). This could lead to a model wherein the TDWG Ontologies, Darwin Core, SDD, and others are (re-) designed to be reference ontologies from which sub-communities robustly build ontologies for their principal uses without social/technical conflict. As an example, it could lead to easing conflicts between the needs of users of Darwin Core for specimen

management and those using it for observations of not-necessarily curated living organisms. See especially the BioOntology Reference Ontologies[169].

- The platform implements several important criteria of the emerging W3C Principles of Good Practice for Managing RDF Vocabularies and OWL Ontologies[170] in part by adhering to the Ontology Metadata Vocabulary (OMV) for describing ontologies and the Protégé Changes and Annotations Ontology[171] (ChAO) for tracking the provenance of ontologies.

Collaboration with other tools

- An extensive REST API[172] allows programmatic access to attributes of an ontology, to terms and relations between them, and to existing views in the sense described above.
- The new NCBO funding provides for the development of a lightweight repository integration, wherein BioPortal installations will be able to update their content from a "master" BioPortal instance, as the content there changes (e.g., new ontology versions are uploaded). No schedule has been set for this facility.

Sustainability

- It is an open source platform with a number of adoptions (some mentioned below), including its driving one, the National Center for Biomedical Ontologies (NCBO) BioPortal installation. NCBO has recently received five years of funding to continue development and support of the BioPortal Platform. Installation and extension of the platform are welcomed by the developers.

. Extensibility

- BioPortal is insulated from the format of the resource, as long as such resource is class-oriented. Six formats are presently supported, including OWL and the OBO format. An OWL2 loader is planned for early in 2011.
- The extensibility and application integration mechanisms of the platform are very ontology-centric. It is being integrated with WebProtégé, a web client version of the Protégé ontology editor. Examples of important applications are: text annotation[173]; browsing heterogeneous resources[174]; web service access to ontologies[175]; and ontology widgets[176] that enable web developers to embed ontology services on HTML pages. Other examples are at the BioOntology Collaboration page[177].
- The current version of the platform is configured by metadata instances of a generic ontology metadata ontology, which furthermore can be accessed via the API of Protégé. The former capability makes it easier for platforms to configure

[169] http://www.bioontology.org/reference-ontologies

[170] http://www.w3.org/2006/07/SWD/Vocab/principles

[171] http://Protégéwiki.stanford.edu/wiki/ChAO_API

[172] http://www.bioontology.org/wiki/index.php/NCBO_REST_services

[173] http://bioportal.bioontology.org/annotator#

[174] http://bioportal.bioontology.org/resources

[175] http://www.bioontology.org/wiki/index.php/NCBO_REST_services

[176] http://www.bioontology.org/wiki/index.php/NCBO_Widgets

[177] http://www.bioontology.org/collaboration

exactly what attributes of ontologies they wish exposed and reasoned upon, and the latter supports developers adding functionality to the platform.

Adoption

- In addition to the mature NCBO BioPortal installation, a BioPortal installation is the test bed[178] of the Open Ontology Repository (OOR). If that becomes the OOR platform, GBIF would be able to deposit all its ontologies in OOR with ease. Even if it does not, the test will establish which OOR high level requirements[179] BioPortal can meet. The DataONE Earth Sciences Semantics Portal (ESSP)[180] is a BioPortal instance installed as a test bed for a node in the DataONE network[181]. The platform has been in use for several years by the Marine Metadata Interoperability Initiative (MMI)[182].

3.1.1.2 Burdens on GBIF for adoption of BioPortal

End user documentation may be sparse pending production of more under the new NCBO grant.

No explicit user training is planned, although it is likely that it will be touched upon at the 3-4 annual Protégé training workshops.

The BioPortal platform is tied in its present implementation to aspects of Protégé APIs. Although this does not limit it to Protégé as an ontology editor, it is likely that Protégé, including WebProtégé, will remain the most widely used editor coupled to the platform. WebProtégé is more convenient as a collaborative ontology development environment than the standalone Protégé, and at least two projects have customised their WebProtégé installations with project-specific user interface (UI). It is possible that GBIF will find that its early adopters would be satisfied with stock WebProtégé, but determine that customizing the UI speeds or eases adoption.

The BioPortal project will set its implementation priorities for the new version based on the needs of its partners and sponsor, which are largely biomedically oriented. GBIF may not have much voice in those priorities.

3.1.1.3 Missing functionality

BioPortal's design does not presently support the lifecycle of flat vocabularies, so GBIF may need to develop mechanisms that allow its flat vocabulary maintenance tools to exchange data with a BioPortal installation, probably by the BioPortal REST APIs. A SKOS editor[183] plugin for Protégé 4 is under development, but it is not presently clear what the nature of its integration with BioPortal may be.

GBIF should evaluate the recommendations of the nearing release ISO 25964[184] standards for vocabulary development as to their applicability to its tool adoption for flat vocabularies (See also Zeng 2009), which briefly discusses the different roles of SKOS

---

[178] http://oor-01f.cim3.net/

[179] http://ontolog.cim3.net/cgi-bin/wiki.pl?OpenOntologyRepository_Requirement

[180] http://d1sweb.dataone.utk.edu/home/release

[181] https://www.dataone.org/

[182] http://mmisw.org/orr/

[183] http://code.google.com/p/skoseditor/

[184] ISO 25964-1 "Thesauri and interoperability with other vocabularies ---—Part 1: Thesauri for information retrieval" http://www.iso.org/iso/catalogue_detail.htm?csnumber=53657.

(primarily for publishing vocabularies in machine readable form) and ISO 25964 (primarily for building and managing vocabularies).

3.1.1.4 Relationship to GBIF Vocabularies Server

The GBIF Vocabularies Server is a ScratchPad-based flat vocabulary server. It can serve as a complementary effort to a BioPortal deployment. To the extent that vocabularies so served become full ontologies, migration to a BioBortal deployment would bring to the user community more power and a larger development community. Such a migration approach would conceive the Vocabularies Server and its underlying ScratchPad (and hence Drupal) as an incubation site for hierarchical vocabularies. Vocabularies for which no requirements are yet identified for the integrative and generative tools of the BioPortal platform would remain maintained in the Vocabularies Server, which would continue to hold what amount to the rationale documents for the more structured vocabularies.

3.1.1.5 Relationship to ISOcat

All three of ISOcat, BioPortal platform, and the GBIF Vocabularies Server provide aspects of vocabulary management, but without study of the ISO 12620:2009 standard—which has some specification of that management—it is difficult to know what the overlap is and whether it would suit GBIF's stakeholders. The latter issue seems of small importance since presently there is very little adoption of formal management systems in the biodiversity community anyway. Of bigger concern is the small size of the development community. We cannot recommend the adoption of ISOcat, though it is worthy of study, particularly for its user interface.

3.1.2 Further tool recommendations

Recommendation 5. Develop a semantically enriched directory of Biodiversity KOS resources by first promulgating a simple ontology of descriptions of such resources (perhaps based on the SPM categories and perhaps as a SKOS application using stock SKOS browsers). An early focus beyond SKOS of this tool should be to exploit some simple relationships of resources or projects to one another (e.g. partnersWith, isComponentOf, usesAsNameAuthority, etc.)

Recommendation 6. Substantial investment should be made in the documentation for every KOS tool or resource that GBIF adopts.

3.1.3 Recommended outreach to GBIF members

Recommendation 7. The GBIFKOS Survey should be kept open and periodic, rigorous, examination be made of its results. Likewise, the GBIF Community pages dedicated to comment on the KOS report should remain open. Finally, GBIF should consider opening a section of the GBIF Community site dedicated to ongoing issues of KOS development as it evolves within GBIF, separately from the comments on this report.

Recommendation 8 Production of simple to use tools for semantic resource development remains an active area of IT research. Existing tools are inappropriate for providers without significant IT support. Consequently, GBIF's outreach should take two forms:

- Develop materials introducing the rudiments of semantic processing. Deliver these as part of other outreach, as a view of what is over the horizon. For example, these could be demonstrations, e.g., of the use of class hierarchies to search for records difficult to find without subsumption computations.

- Develop or adopt, and provide instruction on tools that make it easy for providers to deliver data to semantically rich systems. For example, in collaboration with Morphbank, develop and deliver materials by which providers can easily contribute images to Morphbank and annotate them with morphological terms for use with Morphster.

Recommendation 9. As part of any major vocabulary development effort, consider operating a VoCamp[185] collaboratively with other biodiversity informatics projects, partners and meetings. These should expand the scope beyond focus on occurrence records, and address such concerns as invasive species, endangered species, human impacts, habitat management and others such as GBIF may identify. The VoCamps should follow the guidelines of the VoCamp wiki and should engage combinations of domain scientists and knowledge management specialists.

3.1.4 Recommendation about partnerships

Recommendation 10. Cement partnerships with collaborators in KOS funding proposals now under review with funding agencies. Do not let these relationships lapse if the proposals are not funded or when the projects complete. Become known as a conduit to developers savvy in biodiversity informatics, whether GBIF employees or simply in the GBIF community.

[185] http://vocamp.org/wiki/RunningAVoCamp

4 Next Steps

Both from our own discussions and experiences, as well as some of the messages we take from commentators on the early draft of this report, we know that there are—and will probably always be—two segments of the practitioners that GBIF must engage in any advances in Biodiversity knowledge organisation. The first comprises those whose primary mission is to advance the data-centered parts of biodiversity science. Besides their support of data gathering and curation, these practitioners presently focus mainly on data and metadata solutions for integration and retrieval based on flat vocabularies with formal syntax but informal semantics. The second segment of the community comprises those who already, or intend to, exploit vocabularies that are expressed in languages that enable semantic integration, retrieval, annotation, and reasoning on biodiversity data. These communities overlap, but at the moment their vocabulary lifecycle tools and many of their applications do not. Yet the 25 year history in the biomedical informatics community—evolving from the Unified Medical Language System in 1986—establishes that a rich semantic approach to scientific data can provide for more robust and deeper use of it. Nevertheless, it is clear that even the first segment of the community largely accepts the importance of URIs, and their role in enabling cross-referencing and the reuse of terms. GBIF should aggressively promote their use by its members and data providers, including assisting with issuance and dereferencing, as well as promulgating applicability statements of other communities' URIs for vocabulary terms not defined by GBIF or its partners.

For GBIF to help lead down a path with similar success to that of the biomedical community, it will need to recognise that in early days the two communities of biodiversity informaticists may be professionally somewhat separated but with substantial motivation in the long term to merge, and in the short term cooperate. This means that GBIF must constantly plan how investments in time and funds made in un- or slightly-structured vocabularies are not wasted when the applications they support need to evolve into semantically enriched applications. Some of the low-hanging fruit of simple hierarchies can give immediate and large benefit and do so in a context that is analogous to some current or accelerating biodiversity science and informatics practices. For example, the use of hierarchical classification is familiar in taxonomy and phylogenetics; the importance of rules for the structure of descriptions will be recognised as the same intellectual enterprise as undertaken by the maintainers of the codes of nomenclature; the critical nature of unambiguous ways to specify the names of taxa can ease explanations conveying the importance of URIs and explaining the value of immediate investment in their issuance and use; the utility of subsumption will be almost too obvious to justify (e.g., "if something is true of all instances of a genus, then it is true for every instance of every species in the genus"). All of this means that the social and funding impediments to deep and delicate ontologies need not be impediments to an initial emphasis on the development of KOS which advances "only" lightweight reasoning. But in return for a large initial payoff for highly desired uses such as data discovery and integration, GBIF will encounter an additional burden of "future proofing" its KOS directions. It must adopt and encourage practices and tools that do not foreclose the requirements for more sophisticated reasoning. This means that it must keep in close contact—preferably in collaboration—with research projects such as those we identified in Section 2.2.4. More specifically, it should participate in these efforts with the understanding that ontology development is a central near term goal, even if deeper applications have a less immediate deployment horizon. At the same time, GBIF should begin to develop a section of the GBIF Community site for discussion of desirable deeper reasoning examples, such as data-driven scientific hypothesis exploration, data quality control, broad integration over all observation-based earth sciences, etc. Such a library of problems should be a focus of what GBIF is currently doing about the problems listed,

what related projects are doing, where the "event horizon" for deployment seems to be, etc.

Particularly important will be to seek tools with service or programming interfaces that support tool collaboration and well-specified extension points. At each juncture requiring a flat vocabulary, GBIF must ask itself: what will semantic enrichment of this vocabulary enable (from a top-down point of view: where does a semantic form of it fit in a broader scheme), what will be required for that enabling, and how will GBIF's current tools and development plans contribute to that enrichment. In a sense, GBIF must constantly update this report.

Finally, every outreach activity of GBIF, even if not dedicated to Knowledge Organisation, should offer its audience some insight, even if only with simple analogies, of the utility of rich semantics in helping computers to help people avoid biodiversity miscommunication.

5 References

Buitelaar, Paul, Philipp Cimiano, and Elena Montiel-Ponsoda, eds., Proceedings of the 1st International Workshop on the Multilingual Semantic Web (MSW 2010), Raleigh, North Carolina, USA, April 27th, 2010. CEUR Workshop Proceedings Vol 571. Available at http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-571/

Hodge G. Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. 2000. Available at: http://www.clir.org/pubs/reports/pub91/contents.html.

Levin, Siman A. (Editor in Chief), Encyclopedia of Biodiversity, Elsevier, Inc. 2000

Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., Villa, F. (2007). An Ontology for Describing and Synthesizing Ecological Observation Data. *Ecological Informatics, 2*, 279-296

Mungall et al., "Integrating phenotype ontologies across multiple species", *Genome Biology* 2010, 11:R2 doi:10.1186/gb-2010-11-1-r2)

Shaw et al., "Generating application ontologies from reference ontologies." *In AMIA 2008 Annual Symposium,* Washington, DC, 2008. Available at http://sigpubs.biostr.washington.edu/archive/00000227/01/Shaw_Generating-Application-Ontologies-from-Reference-Ontologies.pdf

Washington, Nicole L., Melissa A. Haendel, Christopher J.

Mungall, Michael Ashburner, Monte Westerfield, and Suzanna E. Lewis. 2009. Linking human diseases to animal models using ontology-based phenotype annotation. PLoS Biology 7, no. 11 (November): e1000247. doi:10.1371/journal.pbio.1000247. http://www.ncbi.nlm.nih.gov/pubmed/19956802.

Zeng, Marcia, "Current status of ISO 25964.1", Presentation at 2009 NKOS Workshop, Available at: http://nkos.slis.kent.edu/2009workshop/MarciaZeng.pdf

6 Appendices

6.1 TDWG Vocabularies

The TWDG vocabularies started on a TDWG site[186] and are migrating to Google Code site[187]. They are divided into many largely (but not entirely) independent files. Some of the vocabularies have garnered greater traction than others. In particular, the vocabularies under the Taxonomy heading have been utilised in several applications, and the other vocabularies tend to be useful to the degree to which they are used by the Taxonomy vocabularies. None of the vocabularies has been approved by the TDWG standards committees.

Although the vocabularies use OWL, their semantics are limited to the widespread use of property domain and range restrictions.

Below is more information about the most commonly used vocabularies, followed by a rough categorisation of the vocabularies.

The Species Profile Model Vocabularies

About thirty categories given as subclasses of a class named *InfoItem*[188] express different concerns of biology, such as *Cytology, MolecularBiology, Ecology, Behaviour*, etc. None presently have properties beyond those of the InfoItem base class, whose properties allow the expression of a few common properties in plain text or controlled vocabularies. These properties support the ability to describe the content of the InfoItem and spatio-temporal or taxonomic contexts in which the content is valid.

TaxonConcept.owl, TaxonName.owl and TaxonRank.owl

These three vocabularies define the TDWG view of taxonomy. A given taxon has a name, but any taxonomic name might mean different things to different agents. Hence, the TaxonConcept vocabulary provides ways to define taxa (using specimens, relations to other taxa, and species descriptions) and the TaxonName vocabulary defines a scientific name - the author, the year, the biological code used, etc

The TDWG vocabularies may be roughly organised as follows:

Humans and human institutions

ContactDetails.rdf
Institution.rdf
InstitutionType.rdf
Person.rdf
Team.rdf
Collection.rdf
CollectionType.rdf
Procedure.rdf

Taxonomy (similar to Darwin Core Taxon class)

[186] http://rs.tdwg.org/ontology/voc/

[187] http://code.google.com/p/tdwg-ontology/

[188] http://rs.tdwg.org/ontology/voc/SPMInfoItems.rdf

TaxonConcept.owl
TaxonConcept.rdf
TaxonName.owl
TaxonName.rdf
Taxonomy.owl
TaxonRank.owl
TaxonRank.rdf
SPMInfoItems.rdf
SpeciesProfileModel.rdf

Occurrence (similar to Darwin Core Occurrence class)

OccurrenceRecord.rdf
OccurrenceStatusTerm.rdf
TaxonOccurrence.rdf
TaxonOccurrenceInteraction.rdf
Specimen.rdf

Time and Space (Similar to Darwin Core Location and Event classes)

CyclicityTerm.rdf
GeographicRegion.rdf

Media

DigitalImage.rdf
PublicationCitation.rdf

Meta-metadata (Foundations)

Common.rdf
TermWithSource.rdf
Base.rdf
Core.rdf

6.2 Potential conflicts of interest.

All of the members of the task group contributing to this report are actively engaged in one or more, but by no means all, of the projects described or recommended for use by GBIF. We therefore particularly welcome opinions that differ from those presented here as well as those that support them.

7 Glossary

ABCD; Access to Biological Collections Data

ALA; Atlas of Living Australia

API; Application Programming Interface

BCI; Biological Collections Index

BCT; Biocomplexity Thesaurus

CBD; Convention on Biological Diversity

CLARIN; Common Language Resources and Technology Infrastructure

CMS; Content Management System

CSV; Comma Separated Value

DwC; Darwin Core

EML; Ecological Metadata Language

EOL; Encyclopedia of Life

EQ; Entity-Quality

EUNIS; European Nature Information System

EnvO; Environment Ontology

GBIF; Global Biodiversity Information Facility

GCMD; (NASA) Global Change Master Directory

GEMET; GEneral Multilingual Environmental Thesaurus

GEO BON; Group On Earth Observations Biodiversity Observation Network

GNA; Global Names Architecture

HAO; Hymenopteran Anatomy Ontology

HCLS; (Semantic Web) Health Care and Life Sciences

HTML; HyperText Markup Language

HTTP; HyperText Transfer Protocol

IETF; Internet Engineering Task Force

IH; Index Herbariorum

IPNI; International Plant Names Index

IRMNG; Interim Register of Marine and Non-marine Genera

ISO; International Organization for Standardization

ITIS; Integrated Taxonomic Information System

IUCN; International Union for the Conservation of Nature

KOS; Knowledge Organisation System

LOD; Linked Open Data

LSID; Life Sciences Identifier

MMI; Marine Metadata Interoperability Initiative

MRTG; Multimedia Resources Metadata Task Group

NBII; National Biological Information Infrastructure
NCBI; National Center for Biotechnology Information
NCBO; National Center for Biomedical Ontology
NGO; Non-Governmental Organisation
NLM; National Library of Medicine
NSF; National Science Foundation
O&M; (OGC) Observations and Measurements
OBO; Open Biological and Biomedical Ontologies
OBOE; Extensible Observation Ontology
OGC; Open Geospatial Consortium
OMV; Ontology Metadata Vocabulary
OSR; (TDWG) Observations and Specimens Records
OWL; Web Ontology Language
PATO; Phenotypic Quality Ontology
RDF; Resource Description Framework
RSS; Really Simple Syndication
SDD; Structured Descriptive Data
SKOS; Simple Knowledge Organization System
SONet; Scientific Observations Network
SPARQL; SPARQL Protocol and RDF Query Language
SPM; Species Profile Model
TDWG; Taxonomic Databases Working Group
TRIN; Taxonomy Research and Information Network
UI; User Interface
URI; Uniform Resource Identifier
WSDL; Web Services Description Language
WoRMS; World Register of Marine Species
XML; Extensible Markup Language