ETHAN: the Evolutionary Trees and Natural History Ontology

Cynthia Sims Parr*[1], J. Sachs[1], A Parafynyk[2] , T.D. Wang[2], R. Espinosa[3],and T. Finin[1]
[1]Department of Computer Science and Electrical Engineering, University of Maryland Baltimore County, USA
[2]Department of Computer Science, University of Maryland, College Park, USA and [3]University of Michigan, USA

1 INTRODUCTION

1.1 Motivation

Large-scale ecological modeling and evolutionary studies often rely on scoring taxon-level characteristics of a wide variety of organisms (e.g. Harvey and Pagel, 1993; Page, 2003; Phillimore et al., 2006; Ives and Godfray, 2006). Compiling such data is laborious and may involve finding and reformatting data tables in original literature, or personally exchanging spreadsheets or ASCII files with researchers. Compiled taxon-level data is beginning to be shared digitally (e.g. Brose et al., 2005) and efforts to support wide data sharing in ecology and evolution (Parr and Cummings, 2005) should make even more compiled data available in forms useful to scientists. However, retrieval, integration, transformation, and validation of shared data in traditional archives remain difficult and largely manual processes. Discovery of new insights from such data is therefore delayed if it is even possible.

Our interest in natural history information stems from our work on a suite of tools to support invasive species biologists. Though food web structure has been recognized to play a role in the success or failure of potential species invasions (reviewed in McCann, 2000), and their impacts (e.g. Thompson, 2005) few ecosystems have been the subjects of empirical food web studies. Thus response teams are typically unable to get quick answers to questions like "what are likely prey and predator species of the invader in the new environment?" We have developed a food web constructor which currently uses an algorithm relying on taxonomic or phylogenetic relationships to model ecological interactions (Parr et al., 2006). Future algorithmic developments will use similarity in life history, natural history, or behavior to inform link predictions.

1.2 Approach

We propose that a semantic web approach using ontologies may address a number of challenges in compiling taxon-level comparative datasets.

In computer science, an ontology is a formal conceptualization of a domain. Typically, it specifies the classes of objects that exist, the relationships among those classes, and the possible relationships among instances of the classes. The W3C's Web Ontology Language (OWL) uses XML both for representing ontologies, and also for making assertions about a given ontology. Each assertion is given in the form of an RDF triple, or subject-predicate-object statement, where parts of the triple reference a term from a specified ontologies or declare values.

A semantic web page begins by listing (as URLs) the locations of the ontologies to be used, then goes on to use those ontologies to make assertions about datasets, human beings, items for sale, etc. A computer agent, on coming to such a page, can ingest the mentioned ontologies and, with the aid of a reasoning engine, use that information to understand the semantics of the ensuing assertions (Berners-Lee et al., 2001).

As an example, consider the "friend of a friend" (FOAF) ontology which is widely used to describe people, their properties (e.g., email addresses) and relationships among them (e.g., who knows whom). The FOAF ontology is described by the URI http://xmlns.com/foaf/0.1/. One can use these terms to embed an assertions in a web page that an individual person with first name Cynthia and last name Parr has the email address csparr@umd.edu. Using the XML encoding those assertions look like the following.

```
1  <rdf:RDF xmlns:foaf="http://xmlns.com/foaf/0.1/">
2  <foaf:Person>
3      <foaf:firstName>Cynthia</foaf:firstName>
4      <foaf:surname>Parr</foaf:surname>
5      <foaf:mbox rdf:resource="mailto:csparr@umd.edu"/>
6  <foaf:Person>
```

Here the first line establishes foaf as shorthand for URI that defines the friend of a friend ontology. Lines 2-6 introduce a new individual person and lines 4-5 specify its properties. A software agent that understands the semantic web language RDF will understand this to mean "there is a person with first name Cynthia and last name Parr who has an email address csparr@umd.edu.

A single semantic web document can both define a class, and also introduce and describe members of that class. Thus, there is not a sharp distinction between "ontology" documents on the one hand, and "instance data" documents on the other. It is helpful, however, to keep this conceptual distinction (between schema and data) in mind. The reader should be aware, therefore, that we refer to "ETHAN" in two distinct senses. When we refer to the OWL documents that define the terms and properties necessary to express natural history and evolutionary tree information we use "ETHAN ontology." When we refer to the taxon accounts we have generated in OWL with information from ADW we call them ETHAN documents.

Below are some of the challenges in compiling taxon-level natural history datasets. We describe how publication of data on the semantic web may address the challenge.

1. **The need to consult many sources because one's scale or scope is rarely the same as that of any one previously published research study.**

   Semantic web documents can be retrieved and collated from all over the World Wide Web by automated processes so they can be assembled according to a user's preference. For example, some reported data may not apply to all populations of a species but a semantic web document can include semantic metadata that characterizes the geographic scope of the data. A semantic web agent can then use reasoning to interpret what falls into the user's scope regardless of the scale of the metadata.

2. **Determining if one's definition of a characteristic matches somebody else's – schema matching**

   Because semantic web documents use ontologies to define terms, a researcher can search for data published using the same terms or semantically related terms. A controlled vocabulary might seem to be a simple solution, but two controlled vocabularies may use the same string to refer to semantically different concepts. Ontologies are rich enough that similar concepts are often identifiable even if they are labeled slightly differently and different concepts can be distinguished even if labeled similarly (Bailin and Truszkowski, 2001).

3. **Resolving conflicts in data when multiple datasets are merged. For example, can we identify when two sources report non-overlapping ranges in body masses for the same species?**

   Ideally, semantic web documents cannot only identify when such conflicts exist (because they are logically inconsistent) but they can help resolve conflicts by automatically assessing trustworthiness of sources based on criteria set by users (Kagal, et al. 2006). In practice, this is difficult and we will discuss limitations.

4. **Not knowing how much data exists for a certain characteristic until considerable time has been invested looking for it.**

   Because agents do much of the work, a semantic web approach can reduce the time necessary for retrieval, transformation, and integration (assuming that the documents are out there!).

5. **Transforming data or making estimates to fill gaps in data requires special knowledge, logic and/or assumptions about related organisms.**

   A semantic web approach is unlikely to totally remove the need for experts to manually review and transform data. However, transformations such as standardization of units can be handled without much user input because ontologies specifying the relationship between units exist. Inheritance and aggregation of characteristics in a concept hierarchy are automatic (inherent to this way of representing knowledge) so "gaps" would be filled using logical inference.

6. **Looking for consistency errors in data is very time consuming.** A prime difficulty is standardizing taxonomy so that reported measures for obsolete names are recognized as applying to current taxon names.

Ontologies that mediate mapping of terms or instances across different sources would be helpful.

7. **Keeping datasets updated is a chore**

   Processes to obtain and merge component datasets are more easily automated if both datasets and processes are maintained and described formally on the WWW.

8. **Sometimes data is made available only in paragraph text**.

   If a well-constructed ontology is available, future researchers or database administrators can use it to make their data available this way. For legacy paragraph text, ontologies can be used by natural language processing systems to harvest meaningful data from the text (e.g. Nirenburg, et al. 2005; Spasic et al. 2005).

For our own work, building the ETHAN ontology is the first step towards a robust process of gathering and maintaining appropriate data for our food web constructor algorithms (Parr, et al. 2006).

1.3 Previous work

The genomic, developmental biology, and biomedical communities have tackled similar problems by creating centralized databases (e.g. GenBank (2006) and more recently by building ontologies, e.g. Gene Ontology (2006), FlyBase (2006), Open Biomedical Ontologies (2006)). However, comparative work at the species level or above in these fields are typically restricted to only model organisms (e.g. Chen et al., 2006, GOMD, 2006). Scientists who want to extend analyses to taxa beyond these well-studied organisms, living and evolving in real populations and communities, will find only scattered digital resources to support them.

Increasingly sophisticated taxonomic name (uBio: MBL/WHOI,2006; ITIS ,2006) and phylogeny (TreeBase, 2006; Tree of Life: Maddison, 2006) resources are now available. Indeed, a similar RDF-based approach has been proposed for the names in TreeBase (Page, 2006). Also, large amounts of ecological and life history data do exist in online databases (e.g,. Myers, 2006, Froese and Pauly, 2006) but they are not standardized with each other nor shared in machine-readable formats across projects.

In the ecological and organismal communities, metadata standards are largely aimed at annotating data collected on individual organisms in particular populations or communities or controlled studies (SEEK, 2004; Bradbury and Clark, 2006) or supporting analysis workflows (Bowers et al., 2005). They are not intended for use for species-level representation. A prototype ontology for the Animal Diversity Web (Parr et al., 2005) emphasized the display and organization of species-level keyword data on that site, and contains no data instances or taxonomic information. The present work replaces the first Animal Diversity Web ontology with a more general framework, ETHAN. We also describe the publication process whereby documents including instance data are published on the World Wide Web.

1.4 Outline of the paper

In this paper, we will describe the Animal Diversity Web and the technologies we chose to use in working with the ADW data. Then we provide an overview of the ontologies and the process used to publish the final semantic documents. We present examples of how the ontology can be used. We present new modeling challenges that arose when creating the ETHAN and how we solved some of them. Finally, our discussion focuses how the ontology might be extended or reused, and on how this solution might scale.

2 MATERIALS AND METHODS

2.1 Animal Diversity Web background

Animal Diversity Web is an online encyclopedia about animals (Kingdom Animalia) originally designed for use in undergraduate biology courses. It has since been used in several comparative studies because it includes information across Kingdom Animalia across the globe. Pages of information, or taxon accounts, are available on over 3000 organisms, mostly at the species level. Coverage is not even, of course, and not all taxon accounts include all possible harvestable data. Most species-level information has been collated by undergraduates from a variety of primary and secondary sources; these are edited by faculty and graduate students. Thus, this is not primary-literature information but a secondary compilation consistent with similar online encyclopedias.

ADW's taxonomic backbone has been constructed from a variety of sources including ITIS (ITIS,2006), Mammal Species of the world (Wilson and Reeder, 1993), the EMBL reptile database (Uetz, 2003), and Howard and Moore for birds (Howard, and Moore, 2003). It includes some phylogenetic information at higher levels in the tree (Parr et al., 2004).

While ADW pages include a large amount of unstructured text, there is increasing use of structured data. Several hundred keywords and numerical data fields are presented as checkboxes to authors using an XML-driven template which has been improved over the years. For publication, HTML documents are generated by XSLT scripts that combine XML data templates with data stored in relational data tables (Parr et al., 2005).

2.2 Ontology design process

We attempted to use OWL-DL as the language of the ETHAN but in the end used OWL-Full. As a W3C recommendation, OWL comes in 3 sublanguages or species: OWL-Lite, OWL-DL, and OWL-Full, ordered in increasing expressivity. Each sublanguage is a superset of the previous one. So a Lite ontology is also a DL ontology, and a DL ontology is also a Full ontology. The three sublanguages differ in what and how the OWL constructs can be used. In OWL-Full, users can use all of the constructs with few restrictions, while in OWL-Lite there are the most syntactic restrictions. These syntactic restrictions guarantee certain levels of computational tractability by limiting semantic expressivity. OWL-Lite corresponds to the description logic SHIF(D) and OWL-DL corresponds to SHOIN(D), for both of which sound and complete decisions exist (W3C, 2006) OWL-Full, on the other hand, is undecidable (Horrocks and Patel-Schneider, 2006). Most of ETHAN stays within the decidable fragment of OWL-DL but for various reasons described below we felt it necessary to use OWL-Full.

For automated publication processes we used a variety of programs in C#, PHP, and XSLT because multiple projects and institutions were involved and it seemed reasonable for each programmer to work in the language most compatible with their system. We attempted to avoid any manual changes to ontologies that would be impossible to automate.

To review and test the core keyword and measures ontology, EthanKeywords.owl, we used both Protégé (Stanford Medical Informatics, 2006) and Swoop (Kalyanpur et al., 2005). OWL species validation was performed using WonderWeb (2006).

3 RESULTS

We construct semantic web documents with taxonomic and natural history data using two core OWL-DL ontologies: **E**volutionary **T**ree" **a**nd "**N**atural **H**istory" concept hierarchies. The combination of relevant information from these two ontologies for a particular taxon is, in a way, an instance document. We call them ETHAN taxon documents.

3.1 Evolutionary Tree: ETHAN taxonomic ontology

Several hundred thousand scientific names of species and higher taxonomic levels are represented in a simple class hierarchy without biological ranks as an OWL document at http://spire.umbc.edu/ontologies/EthanAnimals.owl.

For example, the ontology class Corvus corax is a subclass of Corvus.

```
<owl:Class rdf:ID="Corvus_corax">
      <rdfs:subClassOf
      rdf:resource="http://spire.umbc.edu/ontologies/EthanAnimals.owl#Corvus" />
</owl:Class>
<owl:Class rdf:ID="Corvus">
      <rdfs:subClassOf
      rdf:resource="http://spire.umbc.edu/ontologies/EthanAnimals.owl#Corvidae" />
</owl:Class>
```

This allows us to infer from inheritance, for example, that since Corvus corax is a subclass of Corvus, everything that is true for Corvus may also be true for Corvus corax. Assertions about Corvus corax must be at least a subset of what is known about Corvus. An application can sensibly aggregate for a higher taxonomic level what is known about any of its subclasses This way the information can be propagated as expected among high and low-level taxa.

Importantly, because our lowest taxonomic level is not an instance but a class, there is room for even lower taxa (e.g. subpecies or varieties, of which we currently have very few). It will also be possible to create instances representing actual individuals of our lowest taxonomic level.

These data come from ADW, as described above, which in turns gets its data from a variety of sources. Species names are represented as binomials (cf. ITIS). Here we do not store the classification path to the root, as suggested by Page (2006), because this can be generated on demand and in the eventual ETHAN document (see below). An earlier effort included names beyond kingdom Animalia, as harvested from ITIS, the Integrated Taxonomic Information System. The total size of this OWL document (http://spire.umbc.edu/ontologies/ethan.owl) was 47.8 MB, beyond the capacity of many reasoning engines and difficult to use in any application (see discussion). Also, current codes of nomenclature allow duplicate names across kingdoms. To avoid collisions in class names, we restricted our class hierarchy document to only Animalia which currently produces 38 MB document. Similar ETHAN versions are being made available for other kingdoms.

We created a PHP-driven utility at http://spire.umbc.edu/ont/ethan.php which allows a user to generate parts of this taxonomic ontology of interest to their own work. A user (or intelligent agent) can specify whether all taxonomic siblings, children, and parents are to be included from a starting name. For example, to support demonstration queries about fish eating fish, we generated actinopterygii.owl, which is only 24 MB.

Separating the taxonomy into a different ontology from the rest of ETHAN allows us to replace taxonomic trees when desired. We currently use Animal Diversity Web's taxonomic backbone for immediate utility and maximum consistency (to the extent that ADW's natural history information is internally consistent with its taxonomy). However, we could take advantage of emerging RDF-formatted phylogenetic and taxonomic information available at TreeBase (Page, 2006).

3.2 Natural history: ETHAN keywords ontology

The ETHAN keywords ontology (http://spire.umbc.edu/ontologies/EthanKeywords.owl) defines a set of behavioral and natural history concepts related to taxa as well as relationships among those concepts (Table 1). It covers reproductive and physical description categories, as well as quantitative measures such as body mass, metabolic rates, and life spans. It also introduces a way to describe conservation status of organisms.

As with taxonomy, categorical descriptors such as habitat and life history characteristics (represented as keyword strings by ADW), are also represented as classes. For example:

```
<owl:Class rdf:ID="ThingWithReproductionMatingSystem"/>
<owl:Class rdf:ID="MonogamousThing">
       <rdfs:subClassOf>
              <owl:Class rdf:about="#ThingWithReproductionMatingSystem"/>
       </rdfs:subClassOf>
</owl:Class>
```

Related ontologies (e.g. SEEK. 2004; Raskin, 2006) organize physical and ecological concepts themselves into hierarchies. The classes here facilitate organizing taxa into groups sharing a particular characteristic. We want to say that Corvus corax is-a MonogamousThing, which is different from saying that Monogamy is-a MatingSystem. Corvus corax is-a Desert-livingThing, where other ontologies might indicate that a Desert is-a Habitat.

Numerical measurements are handled differently. An average measurement or range of measurements can only apply to a group of organisms, and not to any of the instances (i.e. individuals) in that group. Nor would it be appropriate to allow inheritance of these values to subclasses of a group. Thus we cannot model these values strictly as datatype properties. Instead, we use annotation properties, which in OWL are associated only with a specified class.

However, measurements have many characteristics in common. Therefore, we specified that each annotation property has the range of a Measurement class with datatype properties describing the units, average, typical low and high values (these may or may not be absolute minimums or maximums).

```
<owl:AnnotationProperty rdf:ID="mass">
       <rdfs:range rdf:resource="#Measurement"/>
```

```
      <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
</owl:AnnotationProperty>

<owl:Class rdf:ID="Measurement"/>

<owl:DatatypeProperty rdf:ID="units">
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
    <rdfs:domain rdf:resource="#Measurement"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="average">
    <rdfs:domain rdf:resource="#Measurement"/>
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#decimal"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="low">
    <rdfs:domain rdf:resource="#Measurement"/>
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#decimal"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="high">
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#decimal"/>
    <rdfs:domain rdf:resource="#Measurement"/>
</owl:DatatypeProperty>
```

As we declare a measurement like “Mass” to be both an AnnotationProperty (which typically does not take an object in an RDF triple) and an ObjectProperty (which does take an object in an RDF triple), we have gone beyond OWL-DL and are using OWL-Full expressivity.

These concepts are organized into meaningful hierarchies. For example, the concept of being an introduced species in the Neotropical region is a subclass not only of “NeotropicalThing” but also “IntroducedThing.” Thus one will be able to query to find all organisms that are neotropical, but also all organisms that have been introduced. Similarly, one may query for all reproductive characteristics of taxa and receive both categorical designations like “SemelparousThing” and measurements such as “NumberOfOffspringPerYear.”

It is also possible further elaborate classes by adding properties that are drawn from existing ontologies describing the world. For example, the “DuneOrDesertLivingThing” class could have a defined property of “hasHabitat” with the value Deserts, which in an earth science ontology may be defined with a certain maximum amount of rainfall. A reasoner could infer that any taxon designated as a subclass or member of “DuneOrDesertLivingThing” could also be inferred to live in places with a certain amount of rainfall. Taxa asserted to be a “PalearcticThing” can take advantage of geospatial or political descriptions of what “Palearctic” means. We have not yet elaborated the ontology in this way because it has not been necessary for our needs.

New concepts can be easily added to the EthanKeywords ontology without causing problems for existing legacy data. Re-organization of concepts need not impact legacy data, as the keyword and measurement hierarchies exist only in the EthanKeywords.owl ontology and not in individual taxon account documents (described below).

3.3 Generating ETHAN taxon documents

Figure 1 illustrates the sequence of events and documents involved in creating ETHAN taxon documents. Animal Diversity Web and Spire projects get their taxonomic information from ITIS and several other sources. As mentioned above, content stored in ADW’s MySQL databases is currently published for the general public as HTML taxon accounts. Taxon-related information can be retrieved from the database in XML format by taxon name using a

PHP utility. Guided by such an XML document, we associate the appropriate keywords and measures in EthanKeywords.owl with the appropriate names from the EthanAnimals.owl hierarchy into a lightweight OWL document labeled with the name of the organism, for example, corvus_corax.owl. Each document includes all of the keywords and measures and the hierarchy of taxonomic groups to which that organism belongs. At this time, we only transform species-level documents.

We make the taxon a subclass not only of its taxonomic or phylogenetic parent, but also of its categorical descriptors. For example, Corvus corax (Northern raven) is not only a subclass of Corvus (in turn a subclass of Corvidae) it is also a subclass of "NearcticThing" and of "ThingWithSexualDimorphismSexesAlike".

```
<owl:Class rdf:ID="Corvus_corax">
      <rdfs:subClassOf
      rdf:resource="http://spire.umbc.edu/ontologies/EthanKeywords.owl#NearcticThing" />
      <rdfs:subClassOf
      rdf:resource="http://spire.umbc.edu/ontologies/EthanKeywords.owl#ThingWithSexualDimorphismSexesAlike"/>
</owl:Class>
```

Measurements applying to this particular taxon are expressed by references to keyword properties in the EthanKeywords.owl document (abbreviated below as "kw"):

```
<owl:Class rdf:ID="Corvus_corax">
      <kw:mass>
            <kw:Measurement rdf:ID="Measurement_Corvus_corax_mass">
            <kw:units
            rdf:datatype="http://www.w3.org/2001/XMLSchema#string">g</kw:units>
            <kw:high
            rdf:datatype="http://www.w3.org/2001/XMLSchema#decimal">1625.00</kw:high>
            <kw:average rdf:datatype="http://www.w3.org/2001/XMLSchema#decimal" />
            <kw:low
            rdf:datatype="http://www.w3.org/2001/XMLSchema#decimal">689.00</kw:low>
            </kw:Measurement>
      </kw:mass>
</owl:Class>
```

We have implemented a "Get Owl" button on each ADW page which runs the transformation script that generates these OWL documents. This allows a casual visitor to look at the OWL for each page, and it also provides a mechanism for semantic search engine crawlers such as Swoogle (Ding, et al. 2004) to regularly generate and index these documents (when the crawler follows the link the scripts are run). As Animal Diversity Web updates its published pages in a weekly procedure, semantic search engines will update their indices on their next visit to these pages.

Nearly 3000 taxon accounts generated as part of our testing are available at http://spire.umbc.edu/ontologies/taxa. Currently the average size of a taxon account file is about 15 KB and the average number of OWL triples included is 250.

3.4 Using ETHAN

Data cleaning example

One of ETHAN's main use cases is to detect inconsistencies in legacy data. As it is also our goal to try to incorporate multiple data sources, this will help prevent original data inconsistencies from propagating. Guided by biologists to identify mutually exclusive characteristics, we are able to take this into account when we convert from the original ADW data into OWL.

For example, when describing animal locomotion, we want to characterize an animal as either a flier or a glider but not both, because suites of adaptations for these modes of locomotion likely differ. In our modeling paradigm, we model these behaviors as classes, and denote them as disjoint. Any animal that flies will be a subclass of "Thing-

That Flies", and any animal that glides will be a subclass of the "ThingThatGlides" class. As we retrieve the locomotion data of an animal from ADW to construct its corresponding OWL file, and if this animal is a subclass of both Flying and Gliding, then the OWL reasoner can detect that such class is unsatisfiable. This tells us that there is inconsistency in the legacy data.

Populating databases for ecology

Figure 2 shows how biologists can use ETHAN and various semantic web agents to populate data for their studies. Data providers make data available in OWL using ETHAN classes or other ontologies (Fig. 2a). Web crawlers at Swoogle (Ding et al., 2004) constantly follow links throughout the web and build an index for all semantic web documents it finds (Fig 2b). A scientist can go to the UMBC TripleShop (Parr et al., 2006), a workshop for semantic web data, and submit a query in an SQL-like language called SPARQL (W3C, 2006). For example, the ecologist may want to build a dataset on the maximum body sizes of predators and prey that are all fish. She constructs a query using appropriate ETHAN terms (such as "Actinopterygii" and "Measurement," "Mass," and "Low") and other ontology terms (such as "Predator" and "Prey"), but does not specify the locations of data to be searched. TripleShop checks the Swoogle index and finds any web documents which contain these terms, e.g. ETHAN documents from the Animal Diversity Web, documents from our food web databases, and ETHAN taxonomic documents that include Actinopterygii. These essentially comprise a very large database of information that should be useful for answering the query. The biologist puts all of these documents (or only those she trusts) in a shopping cart and asks the Triple Shop to apply the original query to these documents., telling the computer to do the logic necessary to know what species count as fish. TripleShop produces a comma delimited file that includes predators and prey that are fish, and their body masses. (Fig. 2c) We can then use this data as input into further analyses.

Discovery through data integration and logic

Using the TripleShop, a biologist may issue queries that involve logical relationships across related databases. For example, what are cases where introduced species are known to be predators on species considered threatened under the United States Endangered Species Act? Using a single query, we can integrate information from distributed documents using different ontologies to answer this question. An ETHAN taxon document includes geographic range keywords, specifying for each geographic region whether the species is introduced or native. In EthanKeywords.owl, we defined a concept called "IntroducedThing" that takes as subclasses all of the relevant geographic range keywords. Taxa that are considered threatened under the United States Endangered Species Act are asserted by Animal Diversity Web to be subclasses of the EthanKeywords.owl term "UsfwsThreatenedThing." We included all taxon documents and the EthanKeywords.owl document in a TripleShop dataset. We add to the dataset over 250 food web studies are available in OWL using an ontology called SpireEcoConcepts. We applied to this dataset a SPARQL query that retrieves those food web links where the predator taxon is an "IntroducedThing" and the prey taxon is a "UsfwsThreatenedThing" The ten results are shown in Table 2; they are surprisingly diverse in both predators and prey. We have not specified that the specific population that eats the threatened taxa must be introduced, just that somewhere in the world the predator has been introduced and somewhere in the world the prey is threatened (e.g. Branta Canadensis has one subspecies that is threatened). Therefore these results illustrate the potential for these predators to impact species where they may be threatened, even if they are not currently a threat. Further, these results show that food web studies are occasionally providing information on species of conservation or invasive species interest. Similar queries on other categories of conservation risk find many more results than we can show here.

4 DISCUSSION AND FUTURE WORK

4.1 New challenges

Big documents vs. small documents

The large size of the complete ETHAN taxonomic documents poses a serious challenge for some semantic reasoning engines and agents, especially those set up to process all the information using only machine memory. Even though many inference engines use dedicated databases during information processing, the amount of data has a very significant impact on the processing time. In an attempt to address those issues, we developed tools for selecting only the relevant triples for a particular query. This way the inference engine can avoid the overhead of analyzing huge amounts of data trying to figure out whether it is relevant to a particular question asked by a user or not. It

should be noted that reasoning about OWL triples requires much more computational resources and time that answering a standard SQL query.

Using our online utility at http://spire.umbc.edu/ont/ethan.php, users can generate a subset of taxonomic data. Future work could involve automating this utility to generate reasoner-friendly subsets that cover the entire taxonomic database.

For the above-mentioned food webs, we created a utility (http://spire.umbc.edu/ont/spirewebs.php) that generates OWL data for the food web studies currently stored in our relational database. Again, instead of storing all the data about all the food webs known to the our project in one single OWL file, a separate data file is created for each food web study with the intent to do some pre-filtering of the data (for example, by habitat if the query deals only with terrestrial taxa) when the query comes in, instead of returning all available data and making the reasoning engine solely responsible for excluding irrelevant information.

OWL modeling issues: implications for SPARQL queries

A common dilemma in ontology design is whether to model information as classes or as individuals. For example, we currently express the fact that gingko-toothed beaked whales live in the Indian Ocean by indicating there is an OWL class, Mesoplodon ginkgodens, which is also a subclass of the class "IndianOceanThing." We could have modeled Mesoplodon ginkgodens as having a certain geographic range keyword as a datatype property.

```
<owl:Class rdf:ID="Mesoplodon_ginkgodens">
       <kw:geographic_range
       rdf:datatype="http://www.w3.org/2001/XMLSchema#string">IndianOceanThing</kw:geog
       raphic_range>
</owl:Class>
```

This would enable us to issue simple SPARQL queries to retrieve all species that live in the Indian Ocean. However, OWL-DL has no provision to map assertions about a class to members, or instances, of the class, so we couldn't say anything about where individual whales live. There is a sense in which what we really mean is

```
<owl:Class rdf:ID="Mesoplodon_ginkgodens">
       <rdfs:subClassOf>
              <owl:Restriction>
              <owl:onProperty rdf:resource="kw:geographic_range"/>
              <owl:hasValue>Indian Ocean</owl:hasValue>
              </owl:Restriction>
       </rdfs:subClassOf>
</owl:Class>
```

This restricts possible ranges for any individual whale that is a member of Mesoplodon gingkodens to the Indian Ocean. However, modeling a measurement like "average body mass" this way is problematic because such restrictions would be inherited. Furthermore SPARQL queries on a particular owl:Restriction are especially complex, Our current method of representation handles this gracefully. Because the class Mesoplodon ginkdodens is a subclass of "IndianOceanThing," any instance of it will also be an instance of "IndianOceanThing." We can subclass multiple classes using additional boolean operators. For example, a taxon Corvus corax can occur in numerous places. We express that by saying that Corvus corax is a subclass of the union of several geographic range classes, for example, "NearcticThing" OR "PelearcticThing."

This means any individual of Corvus corax is assumed to be instances of these two geographic designations. However, if a user attempts to assert that an individual of Corvus corax is from Indian Ocean (instance of "IndianOceanThing"), then a logical conflict will result, since "IndianOceanThing" is disjoint from either "NearcticThing" or "PelearcticThing" in ETHAN.

Top-down vs. bottom up approach

Some data authors may only provide data at the species level, while others may not have this information available to them. Currently we only publish keywords in species-level ETHAN documents because this is the level of infor-

mation provided by ADW. This bottom-up approach allows an application to use information from species to make statements about higher levels in the taxonomic hierarchy, but does not take advantage of inheritance. A characteristic common to all members of a family is asserted for every species in that family.

However, the ETHAN ontology framework can handle inheritance gracefully just as illustrated in the example above. As long as an assertion about a class is logically consistent with what is represented for classes above and below it in the taxonomic hierarchy the ontological framework is unconcerned about where the information is asserted. We suspect, however, that biologists would require an application displaying data using the ETHAN ontology to clearly identify which knowledge was directly asserted by authors and which was inferred by inheritance or aggregation. In fact, ADW does have information asserted at other levels and in future work we can incorporate that into ETHAN documents.

Limitations

Our use of annotation properties for declaring the values of measurements presents some difficulty for current editing applications, such as Swoop (Kalyanpur et al., 2005), Protégé (Stanford Medical Informatics, 2006), and Photostuff (Halaschek-Wiener, et al. 2005). First, values for the annotation properties' datatype properties are not displayed and can therefore not be edited using Swoop or Photostuff. However there appears to be nothing wrong with the legality of the code (which we generate using programs) because reasoners and validators behave as expected and we are not to our knowledge violating rules of OWL. Second, Swoop identifies the species of OWL being used by ETHAN documents as OWL-DL while Protégé appears unable to determine the species (though documents pass all tests for OWL-DL). These problems are probably not inherent to ETHAN but to the implementation of the above-mentioned programs. To be fair, Swoop and Photostuff are a lightweight editors not intended for OWL-Full editing, and Photostuff is designed primarily for instance editing.

An additional concern is that OWL-DL does not permit the declaration of a domain or range for an annotation property. Yet it would make sense to restrict use of the "Wingspan" annotation property to only those classes (e.g. Aves or Insecta) that actually have wings. OWL-Full does allow this, but we did not wish to further complicate the ability to reason.

Another serious issue is that by modeling measurements this way we cannot check measurements using OWL reasoning. Instances of annotation properties cannot have a logical relationship to each other, so we cannot test whether measurements are consistent across sources (for example, use generic reasoners to determine that a report value for a species falls outside a range of reported values for a genus). However, it is possible for an external application to extract measurements about organisms and perform such consistency checks.

4.2 Extending and scaling

Other than serving as a general model, ETHAN will be a significant contribution if others can use it or extend it, and if it scales well for large databases.

First, must everyone wishing to share species-level data use ETHAN? The answer is no. While it may be useful for projects that are just beginning to design their schema and who have not yet gathered data, existing projects may want to develop their own ontology using their own class names and hierarchies. Then the problem becomes one of creating a mapping from one ontology to others. The mappings become part of the semantic web and are available to others who wish to combine instances from those ontologies. In fact, we have not taken advantage of existing ontologies for measurement (e.g. Raskin, 2006; SEEK 2004) because it was easier for our developers to keep ETHAN self-contained and not import largely irrelevant external ontologies. However in future work we can create equivalencies from our measurements to that of others.

Second, how reusable is ETHAN? We anticipate that it can be used for a wide variety of legacy websites that currently make their data available only in XML or HTML. For example, FishBase (Froese and Pauly, 2006) contains many of the same kinds of information as ADW and it would be interesting to see how consistent data from the two sites are. By mapping the FishBase schema to the ETHAN ontology and following a similar semantic web document generation process it should be possible to rapidly generate OWL documents whose information can be merged with ADW and checked for consistency.

Third, how extensible is ETHAN? If database managers expect considerable overlap with ETHAN but will be covering some areas in more detail, ETHAN can be extended. This is what we are planning to do for the LepTree project (http://www.leptree.net), which is storing information about Lepidoptera taxa. This project's focus is more on morphology and evolution than Animal Diversity Web's, so LepTree's ontology will extend ETHAN by adding many more detailed classes in these areas.

Finally, how well does this technology scale? Animal Diversity Web is already designed to cover a global scale across all animals. ADW is a constantly growing resource. Adding instance data should not pose significant challenges for ETHAN because each ETHAN taxon document is small; as mentioned above; this should assist efforts to keep reasoning problems under control. We expect different ontologies to represent knowledge about non-animals to be needed.

There has been skepticism about the success of an OWL-based semantic web approach (Festa, 2005). Widespread adoption of these techniques has been slow in coming and some argue that more lightweight approaches such as microformats can solve many data integration problems (Khare and Celik, 2006). Indeed, the implementation of our effort was quite time consuming. However, consider that for large scientific datasets, current database approaches can handle easy tasks on large amounts of data. Semantic web approaches have so far enabled complex tasks only on small amounts of data. Our work represents an attempt to make complex tasks possible on large amounts of data.

5 CONCLUSIONS

We present ETHAN, a framework for publishing taxon-level natural history characteristics on the semantic web. We show how information on over 3000 taxa from Animal Diversity Web (http://www.animaldiversity.org) is made available in this way. The availability of such data in OWL format makes machine-assisted integration and querying of this information possible. Some challenges are addressed in this work, while others remain for future developers.

6 REFERENCES

Bailin, S. C. and Truszkowski, W., 2001: Ontology Negotiation between Scientific Archives. . Proceedings of the 13th International Conference on Scientific and Statistical Database Management, July 18-20, 2001, George Mason University, Fairfax, Virginia, USA, Pp. 245-250.

Berners-Lee, T., Hendler, J. and Lassila, O., 2001. The semantic web. Scientific American 284, 28-37.

Bowers, S., Thau D., Williams R. and Ludäscher B., 2005. Data procurement for enabling scientific workflows: On exploring inter-ant parasitism. Proceedings of the 2nd International Workshop on Semantic Web and Databases (SWDB), Lecture Notes in Computer Science v. 3372, p. 57-63. Springer Verlag,

Bradbury, J. and Clark, A., 2006. Animal Behavior Ontology. Available at: ethodata.org.

Brose, U., Cushing, L., Berlow, E.L., Jonsson, T., Banasek-Richter, C., Bersier, L.-F., Blanchard, J.L., Brey, T., Carpenter, S.R., Blandenier, M.-F.C., Cohen, J.E., Dawah, H.A., Dell, T., Edwards, F., Harper-Smith, S.J.U., Knapp, R., Ledger, A.M.E., Memmott, J., Mintenbeck, K., Pinnegar, J.K., Rall, B.C., Rayner, T., Ruess, L., Ulrich, W., Warren, P., Williams, R.J., Woodward, G., Yodzis, P. and Martinez, N.D., 2005 Body sizes of consumers and their resources. Ecology 86, 2545

Chen, Y.-C., Hsiao, C.-D., Lin, W.-D., Hu, C.-M., Hwang, P.-P. and Ho, J.-M., 2006. ZooDDD: a cross-species database for digital differential display analysis. Bioinformatics 22, 2180-2182.

Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Peng, Y., Reddivari, P., Doshi, V. and Sachs, J., 2004. Swoogle: a search and metadata engine for the semantic web. Proceedings of the thirteenth ACM international conference on Information and knowledge management p. 652-659. ACM Press, New York.

Ecological Metadata Language (EML). Available at: http://knb.ecoinformatics.org/software/eml/.

Festa, P. 2005. Next big step for the Web--or a detour? CNETNews.com. Available at: http://www.zdnetasia.com/news/hardware/0,39042972,39220930,00.htm

FlyBase. FlyBase: A Database of the Drosophila Genome. Available at: http://www.flybase.org/.

Froese, R. and Pauly, D., 2006. FishBase. Available at: http://www.fishbase.org.

GenBank. Available at: http://www.ncbi.nlm.nih.gov/Genbank/.

Gene Ontology. Available at: http://www.geneontology.org.

Generic Organism Model Database (GOMD) project. Available at: http://www.gmod.org/.

Halaschek-Wiener, C, Schain, A, Golbeck, J., Grove, M., Parsia, B. and Hendler, J., 2005. A flexible approach for managing digital images on the semantic web in 5th International Workshop on Knowledge Markup and Semantic Annotation (Galway, Ireland). Available at http://www.mindswap.org/~chris/publications/PhotoStuffSemannot2005.pdf

Horrocks, I. and Patel-Schneider, P.F., 2003. Reducing {OWL} Entailment to Description Logic Satisfiability. In: Fensel et al. (Eds) Proc. of the 2003 International Semantic Web Conference (ISWC 2003), Lecture Notes in Computer Science, No 2870; 17-29. Springer.

Howard, R. and Moore, A., 2003. A complete checklist of the birds of the world, 3rd Edition. Princeton University Press, Princeton, NJ.

Khare, R. and Celik, T., 2006. Microformats: a pragmatic path to the semantic web. In: Proceedings of the 15th International Conference on World Wide Web, Edinburgh, Scotland; 865-866.

ITIS. 2006. Integrated Taxonomic Information System (ITIS). Available at: http://www.itis.usda.gov.

Ives, A.R. and Godfray, H.C.J., 2006. Phylogenetic analysis of trophic associations. American Naturalist 168:E1-E14.

Kalyanpur, A., Parsia, B., Sirin, E., Cuenca-Grau, B. and Hendler, J., 2005. Swoop - a web ontology editing browser/ Journal of Web Semantics 1(4). Version 2.3 Available at: http://www.mindswap.org/2004/SWOOP/.

Maddison, D.R., 2006. Tree of Life. Available at: http://www.tolweb.org.

MBL/WHOI, 2006. uBio: Universal Biological Indexer and Organizer. Available at: http://www.ubio.org/.

McCann, K.S., 2000. The diversity-stability debate. Nature 405:228-233.

Myers, P., 2006. Animal Diversity Web. Available at: http://www.animaldiversity.org.

Nirenburg, S., McShane, M. Zabludowski, M., Beale, S. and Pfeifer, C., 2005. Ontological semantic text processing in the biomedical domain. Working Paper #03-05, Institute for Language and Information Technologies, University of Maryland Baltimore County.

Open Biomedical Ontologies. Available at: http://obo.sourceforge.net/.

Page, R.D.M., 2006. Taxonomic names, metadata, and the semantic web. Biodiversity Informatics, 3, 1-15.

Parr, C.S. and Cummings, M.C., 2005. Data-sharing in ecology and evolution. Trends Ecol. Evol. 20, 362-363.

Parr, C.S., Parafiynyk, A, Sachs, J., Pan, R., Han, L., Li, D., Finin, T., Wang, T.D. and Hollander. A., 2006. Using the semantic web to integrate ecoinformatics resources. In: Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06), July 2006; 1949-1950.

Parr, C.S., Espinosa, R., Dewey, T., Hammond, G. and Myers, P., 2005. Building a biodiversity content management system for science, education and outreach. Data Science Journal 4, 1-11.

Parr, C.S., Lee, B., Campbell, D. and Bederson, B., 2004. Tree visualizations for taxonomies and phylogenies. Bioinformatics 20, 2997-3004.

Phillimore, A., Freckleton, R.P., Orme, C.D.L., Owens, I.P.F., 2006. Ecology predicts large-scale patterns of phylogenetic diversification in birds. American Naturalist 168(2):220-229.

Raskin, R. 2006. Guide to SWEET ontologies. Available at: http://sweet.jpl.nasa.gov/. Accessed 18 October 2006.

SEEK Knowledge Representation Working Group. 2004. SEEK ontologies. Available at: http://cvs.ecoinformatics.org/cvs/cvsweb.cgi/seek.

Species2000. Available at: http://www.sp2000.org/.

Spasic, I., Ananiadou, S., McNaught, J. and Kumar, A., 2005. Text mining and ontologies in biomedicine: Making sense of raw data. Briefings in Bioinformatics 6(3):239-251.

Stanford Medical Informatics. 2006. Protégé 3.1. Available http://protege.stanford.edu/.

Thompson, J.K., 2005, One estuary, one invasion, two responses--Phytoplankton and benthic community dynamics determine the effect of an estuarine invasive suspension-feeder. In: R.F. Dame and S. Olenin (Eds.), The Comparative Roles of Suspension-feeders in Ecosystems, Springer Press, The Netherlands; 291-316.

TreeBASE. Available at: http://www.treebase.org/treebase.

Uetz,P. (ed.), 2003. European Molecular Biology Laboratory Reptile Database. http://www.reptile-database.org.

Wilson, D.E. and Reeder, D.M., 1993. Mammal Species of the World. Smithsonian Institution Press, Washington, DC.

WonderWeb. 2006. Available at: http://phoebus.cs.man.ac.uk:9999/OWL/Validator.

W3C. 2006. SPARQL Query Language for RDF. http://www.w3.org/TR/rdf-sparql-query/.

W3C. 2006. Web Ontology Language. (OWL) http://www.w3.org/2004/OWL/.

Figure captions

Figure 1. ETHAN generation workflow. Animal Diversity Web and SPIRE projects get their taxonomic information from ITIS and several other sources. As mentioned above, content stored in ADW's MySQL databases is currently published as HTML taxon accounts. We've added additional processes that publish the keyword ontology that ADW uses, which, when combined with the actual data instances and the taxonomic information, result in an OWL version of the Animal Diversity Web account.

Figure 2. Flow of natural and life history data on the semantic web. a. Database managers use ETHAN terms and taxonomic data to transform their native stores of information into OWL documents, linking these transformations to HTML documents or placing them in web-accessible directories where they can be discovered b. An agent such as Swoogle crawls these semantic web documents and indexes them. Users of Swoogle services can then issue queries to find pointers to documents with terms from ETHAN, regardless of their source or location. c. Using an application like TripleShop, a user can construct a complex query, discover documents to build a dataset, apply the query to the dataset and have results returned in a spreadsheet-compatible format. Queries and datasets can be saved for future use and shared with collaborators.

Table 1. ETHAN Keywords concepts. Concepts are described informally in this table but in ETHAN have specialized labels such as "RainforestThing." Updated from Parr et al., 2005.

Class	Subclasses	Annotation Properties
Geographic range	Nearctic, Neotropical, Antarctica, Indian Ocean, Mediterranean Sea, island endemic, cosmopolitan	
Habitat	temperate, tropical, polar, terrestrial, saltwater/marine, freshwater, desert, rainforest, pelagic, rivers and streams, urban, intertidal	elevation, depth
Physical description	ectothermy/endothermy, type of symmetry, sexual dimorphism, polymorphism, poisonous/venomous	mass, length, basal metabolic rate
Development	neotenic/paedomorphic, metamorphosis, colonial growth, indeterminate growth	
Reproduction: mating systems	monogamous, polygamous, eusocial, cooperative breeding	
Reproduction: general behavior	semelparity/iteroparity, seasonal/year round breeding, gonochoric, hermaphroditic, parthenogenic, sexual/asexual, internal/external fertilization, oviparous/viviparous	breeding season, number offspring, time to hatching, age at maturity
Reproduction: parental investment	presence of parental care, types of parental investment by males and females, altricial/precocial, extended period of juvenile learning	
Lifespan/longevity		expected and maximum lifespan in captivity and in the wild
Behavior	degree of sociality, diurnal/nocturnal, migration, mode of locomotion or dominant way of living (scansorial, fossorial, natatorial), sessile/motile, hibernation/aestivation	territory and home range size
Communication/Perception	visual, chemical, tactile, acoustic, electrical, magnetic, heat, ultrasound, bioluminescence, mimicry, scent marking, pheromones	
Food habits	dominant food type (carnivore, herbivore, other) along with a more specific designation (molluscivore, scavenger, nectarivore, coprophage), list of all foods eaten, special food behaviors including caching and filter feeding	
Predation	mimicry, crypsis, aposematism	
Ecosystem roles	seed dispersal, pollination, biodegradation, soil aeration, creates habitat, keystone species	
Economic Importance for humans: positive	pet trade, food, research, ecotourism, medicine, pollinates crops, controls pests	
Economic Importance for humans: negative	injures humans, crop pest, household pest, causes or carries domestic animal disease	

Conservation status	status on IUCN Redlist and U.S. E.S.A., CITES category	

Table 2. Introduced predators that eat threatened species. Results of a query over multiple semantic web documents for any taxon considered 'introduced" that has prey with "Threatened" status under the USFWS Endangered Species Act. Predators may not have introduced status in the same community where the prey has Threatened status.

Predator (introduced)	Prey (USFWS Threatened)
Canis lupus familiaris	Branta canadensis
Canis lupus familiaris	Cyclura cornuta
Eira barbara	Leontopithecus caissara
Felis silvestris	Plecotus rafinesquii
Mephitis mephitis	Branta canadensis
Mustela frenata	Spermophilus brunneus
Procyon lotor	Plecotus rafinesquii
Salvelinus namaycush	Salvelinus confluentus
Sus scrofa	Cyclura cornuta
Taxidea taxus	Spermophilus brunneus

Figure 1

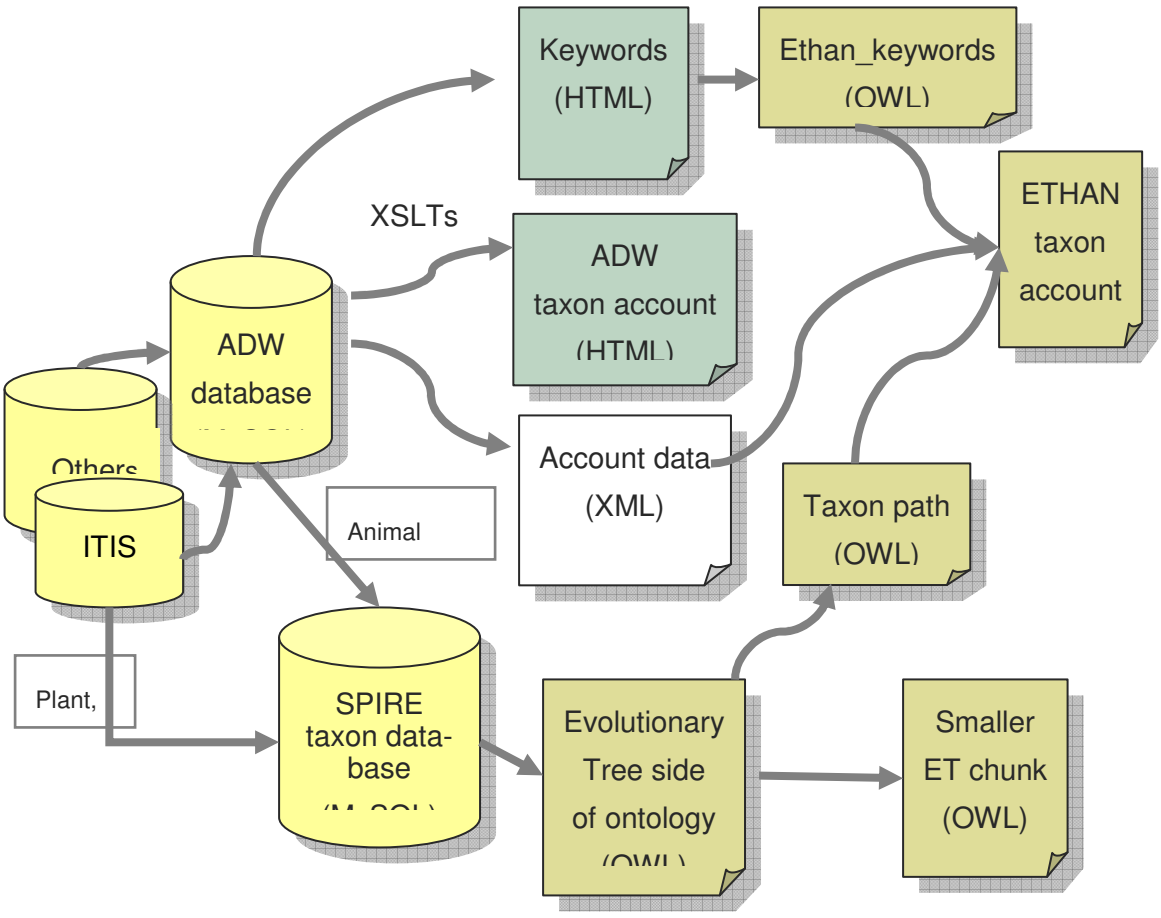

Figure 2