Proceedings of the Sixteenth International Semantic Web Conference (ISWC), Vienna, Austria, 2017.

A Controlled Crowdsourcing Approach for Practical Ontology Extensions and Metadata Annotations

Yolanda Gil[1], Daniel Garijo[1], Varun Ratnakar[1], Deborah Khider[2], Julien Emile-Geay[2], Nicholas McKay[3]

[1] Information Sciences Institute, University of Southern California
{gil, dgarijo, varunr}@isi.edu
[2] Department of Earth Sciences, University of Southern California
{khider, julieneg}@usc.edu
[3] School of Earth Sciences and Environmental Sustainability, Northern Arizona University
Nicholas.McKay@nau.edu

Abstract Traditional approaches to ontology development have a large lapse between the time when a user using the ontology has found a need to extend it and the time when it does get extended. For scientists, this delay can be weeks or months and can be a significant barrier for adoption. We present a new approach to ontology development and data annotation enabling users to add new metadata properties on the fly as they describe their datasets, creating terms that can be immediately adopted by others and eventually become standardized. This approach combines a traditional, consensus-based approach to ontology development, and a crowdsourced approach where expert users (the crowd) can dynamically add terms as needed to support their work. We have implemented this approach as a socio-technical system that includes: 1) a crowdsourcing platform to support metadata annotation and addition of new terms, 2) a range of social editorial processes to make standardization decisions for those new terms, and 3) a framework for ontology revision and updates to the metadata created with the previous version of the ontology. We present a prototype implementation for the paleoclimate community, the Linked Earth Framework, currently containing 700 datasets and engaging over 50 active contributors. Users exploit the platform to do science while extending the metadata vocabulary, thereby producing useful and practical metadata.

Keywords: Metadata, crowdsourcing, semantic wiki, collaborative ontology engineering, semantic science, incremental vocabulary development.

1 Introduction

Existing frameworks for collaborative ontology development assume a clearly phased separation between ontology creation, release of the ontology, and use of the ontology [12, 8, 25]. These frameworks do not fit many areas of science, notably field-based sciences like ecology, Earth, and environmental sciences. These areas are extremely diverse, with data collected by many individual scientists each with idiosyncratic

instruments, methodologies, representations, and requirements. As soon as an ontology is put to the test through practical use, we can anticipate the need for many additions and extensions to accommodate that diversity. Therefore, an ontology would need to be part of a framework that supports constant change while being used. Moreover, the involvement of diverse experts in the community would be needed, raising challenges about incentives and more importantly about coordination of requirements.

Our goal is to support the paleoclimate community, which studies past climate based on the imprint on various systems like trees, glacier ice, or lake sediments. This community employs such a diverse array of data collection and analytical techniques that it has been very challenging to develop shared ontologies. As a result, it is hard to find and aggregate datasets contributed by diverse scientists to paint a global picture of past climate change. Many other scientific communities face similar challenges, as do any organizations with highly heterogeneous or dynamic knowledge environments.

We propose a new approach to ontology development based on *controlled crowdsourcing*. Users (the crowd, who are experts in the domain rather than generic workers) concurrently create new terms as needed to describe their data, making the terms immediately available to others. Once the new terms are agreed upon, they can become part of the next version of the ontology. To coordinate the growth of the ontology and to create necessary incentives, we organize the community so that proper editorial control is exercised. We have implemented this approach in the Linked Earth Framework and deployed it for the paleoclimate community. The Linked Earth Framework is a socio-technical system that includes a crowdsource annotation platform to create metadata and propose new terms as needed, editorial processes to make standardization decisions for new terms and adding them to a core ontology, and a framework for ontology revision and updates. The system contains about 700 datasets and over 50 active contributors organized into 12 working groups.

The paper is structured as follows. Section 2 describes the challenges through a motivating scenario for the field sciences. Section 3 describes our new approach for controlled crowdsourcing of metadata and ontology extensions, followed by a description of our implementation in the Linked Earth Framework in Section 4. Section 5 describes the uptake by the paleoclimate community. The paper concludes with related work and a discussion of current limitations and plans for future work.

2 Motivation: Metadata Diversity in Ecology and Environmental Sciences

Data integration is particularly challenging in ecology and environmental sciences, where data are collected piecemeal by individual investigators with idiosyncratic organization and notation. This makes it very hard to create standards, in contrast with other sciences where there is more uniformity in the collection process such as genomics and astrophysics. Our focus is paleoclimatology, whose goal is to reconstruct the climate in past times based on indicators such as the chemical composition of glacier ice or the width and density of tree rings. These indicators, called *climate proxies*, are obtained from various physical samples including ocean and lake sedi-

ments, ice, cave deposits, corals, and wood. There are many kinds of physical samples, hundreds of types of measurements that can be obtained from them, and hundreds of approaches to use those measurements to reconstruct *climate variables* such as temperature or rainfall. Only by integrating all this incredibly diverse data at planetary scales can we develop an understanding of climate evolution across space and time. Global climate reconstruction efforts are very valuable, for example the Past Global Changes (PAGES) 2k worldwide collaboration, devoted to the study of the climate in the last 2,000 years, published the most cited paper to date in Nature Geoscience [17]. Yet such studies require significant manual integration, and use only a fraction of the available data which is very hard to find and aggregate (e.g., [18]).

Tackling such diversity is a formidable task. The community has developed some basic standards, such as Pangaea's interoperability scheme [19] and the common variable properties in the Linked PaleoData (LiPD) format [14]. These provide a strong substrate for building community databases and integration efforts, but most scientifically relevant properties of the data still have to be found through text searches. Separately collected datasets still have to be aggregated painstakingly by hand. Additional standardization could be pursued, but it would require involving hundreds of scientists in diverse areas who study all types of samples, measurements, reconstruction methods, and variables. For example, a scientist who studies corals would compare isotopic variations among different coral species, whereas a glaciologist would care about the consistency of the methods to measure water isotopes and the composition of ancient gas in air bubbles trapped in ice. Creating standards for such a diversity of data is daunting, and yet crucial for our understanding of past climate fluctuations [4].

The metadata diversity in ecology and environmental sciences poses new requirements to support scientific metadata standardization:

1. **The creation of metadata properties should always be open to new contributors**, so that any scientist is able to suggest new properties based on their expertise and the kinds of samples and measurements that they study.
2. **A community repository should have frequently updated metadata standards**, so that it always offers users the most recent extensions.
3. **Community engagement is crucial.** There must be mechanisms for scientists to see value in the metadata properties being created in terms of enabling them to do research, that is, to find, aggregate, or analyze datasets.
4. **New metadata properties created by contributors must be coordinated and integrated with existing ones in a principled manner.** Any new distinctions that a scientist wishes to make must be related to other proposals, and must be considered in the context of the emerging standard.

There are many important challenges. Could scientists extend a metadata ontology on the fly without receiving training in ontology engineering? Could an ever-evolving ontology be used to annotate data when those annotations will need to be updated in future versions? How can we engage a scientific community that has little interest in participating in standards development (though there may be a strong interest in seeing a standard emerge and using it in their work)? How could we ensure that the extensions proposed by dozens of contributors are principled and capture the complexities involved in sophisticated scientific data, and are useful to do science? How could we avoid redundancies and inconsistencies as new metadata properties are added?

These challenges involve important aspects of community engagement and incentives in addition to technical aspects of usability, collaboration, and knowledge management.

Although these requirements and challenges are motivated by our work in scientific metadata, they arise more broadly in organizations outside of science with diverse and rapidly evolving knowledge.

3 A Socio-Technical Approach for Controlled Crowdsourcing of Ontology Extensions and Metadata Annotations

We propose a novel approach that combines social and technical elements for controlled crowdsourcing of ontology extensions and metadata annotation. Note that crowdsourcing in our case means a large amount of users who are domain experts [16], rather than untrained workers. Our approach has several important features:

- A *metadata crowdsourcing platform*, where any registered user can add metadata for a dataset. In doing so, users can easily create new metadata properties as needed. These activities would be done in the platform as follows:
    - Users are first guided to choose among existing metadata properties. If none suits their needs, then they are invited to create new ones.
    - The creation of a new metadata property is as simple as adding a term and some documentation for it, so that any newcomer can easily do it and it does not get in the way of their work.
    - New properties are created because they are needed by a user to describe the dataset that they are working with. That creates a context and justification for the new metadata property, and in that sense they are practical and useful metadata.
- A *controlled standardization process*, consisting of a social structure of editor roles and working groups. The standardization process manages the community activities to consider the addition to the ontology of new properties suggested by the crowd. This way we ensure that the standard metadata ontology grows in a consistent and principled manner while being driven by the practical needs of the community. The process includes:
    - Quick turnaround decisions for clearly useful and uncontroversial new properties to be incorporated into the standard.
    - Mechanisms for escalating discussions and facilitating decision making for integrating new properties with the existing ontology.
    - Explicit reporting of changes to the ontology in new versions.
    - An initial standard upper ontology that is designed with solid principles, to be extended through crowdsourcing with more specific terms.
- A *metadata catalog evolution framework*, which supports the community to describe datasets using the ontology, use new properties as they are added, and easily transition to new versions of the ontology as they become available.
    - The metadata properties that are part of the standard must co-exist with the ones that are newly created by the crowd.
    - Metadata annotations should be updated with new ontology releases.

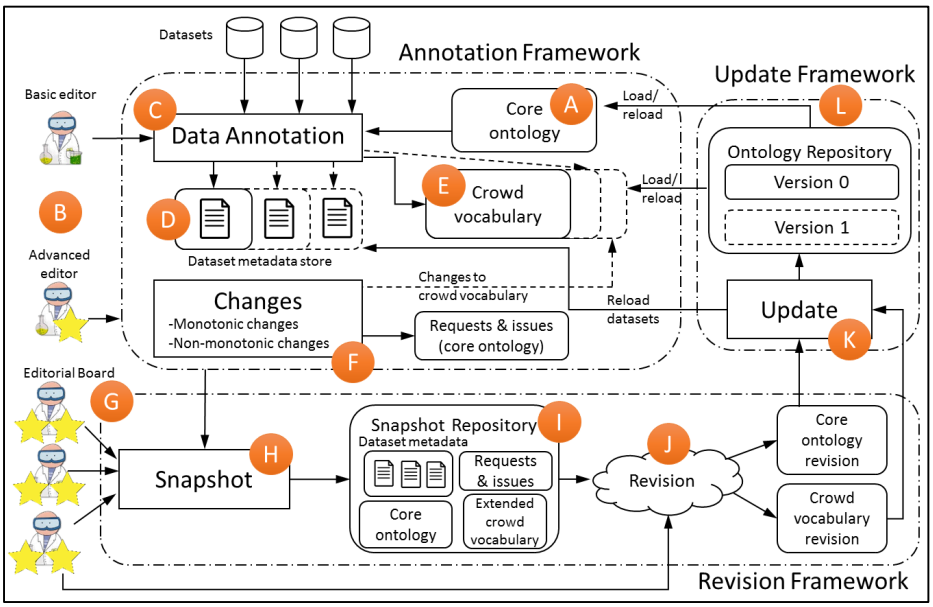

Figure 1: Overview of our approach for controlled crowdsourcing.

Figure 1 highlights the main aspects of the proposed controlled crowdsourcing process. There are three major components of the process: 1) annotation and vocabulary crowdsourcing, shown at the top; 2) editorial revision for ontology extensions, shown at the bottom; and 3) updates to the metadata repository, shown on the right.

An Annotation Framework supports the annotation and crowdsourcing process, shown in the top left of Figure 1. The framework is initialized with a *core ontology* (A). The core ontology represents a standard that the community has agreed to use. Users, which form the crowd (B), interact with a metadata annotation system (C) to select terms from the core ontology for annotating datasets (D). For example, a term such as "archive type" from the core ontology could be used to express that a dataset contains coral. If a term they want to specify is missing, users may propose extensions by simply adding the term, which becomes a property in the *crowd vocabulary* (E), That new term is immediately available to other users when they are annotating their datasets. Users may also have requests and issues (F), such as requests for changes to the core ontology, comments to discuss new proposed terms, and other issues. There are two main types of changes to the core ontology that users may request:

a) *Monotonic changes*: these are proposed new terms for the core ontology that do not affect any prior metadata descriptions made by others. For example, a user may extend the ontology for coral and add the property "species name".
b) *Non-monotonic changes*: these concern existing terms in the core ontology or the crowd vocabulary that have already been used to describe datasets. For example, a request to rename an existing property or change its domain/range.

A Revision Framework supports the ontology revision process, shown at the bottom of Figure 1. A select group of users form an *editorial board* (G) that is continu-

ously reviewing the requests to extend or change the core ontology with crowd vocabulary terms. The editorial board discusses these proposals and determines if they should be incorporated into the standard, effectively beginning to plan extensions to the core ontology. Eventually the editorial board decides to incorporate a new version of the core ontology in the framework. At that point, the editorial board would generate a snapshot (H) of the contents of the platform, which would include the metadata annotations to all the datasets, the crowd vocabulary, and the proposed requests and issues (I). The board would then produce a revision of the core ontology by incorporating terms from the crowd vocabulary where there is agreement (J). This may involve resolving inconsistencies and restructuring the core ontology if there are deeper issues. The crowd vocabulary is also updated to include only the remainder terms.

An Update Framework upgrades the ontologies of the Annotation Framework. Given new versions of the core ontology and the crowd vocabulary, the editorial board updates the metadata annotations of the datasets to reflect those changes (K). This can be done semi-automatically, so that monotonic changes are automated as are some of the simpler non-monotonic changes. More complex changes may need to be done manually. Care must be taken to document all these changes in the "Talk" pages of the wiki, so that users can understand why the annotations they made to their datasets are now done differently. Finally, the Annotation Framework is reinitialized by loading the new versions of the core ontology and the crowd vocabulary and the new versions of the annotations to the datasets (L). The process continues with subsequent waves of crowdsourcing annotations and terms followed by core ontology updates. Note that the editorial board may do revisions as often as they wish, and to postpone consideration of some changes until more information is obtained from the crowd.

Appropriate community engagement is a non-trivial aspect of this approach. The metadata annotation interface must be easy to use by an average user in order to keep them involved. Users must see immediate reward for their annotations in order to continue to be engaged. Editorial board members must be selected so they are representative of the different expertise areas and able to understand the broader implications of each extension. Decisions about the standard must incorporate broad community input to be accepted and adopted in practice. The overall process must be transparent and inclusive so it is trusted as a community effort. Therefore, to implement this approach we must consider a socio-technical system that addresses both community and the technology aspects. The next section describes our work with a scientific community to investigate our approach.

4 The Linked Earth Framework

We are using our controlled crowdsourcing approach in the Linked Earth project to support the paleoclimatology community. As described earlier, the variety of samples, measurements, and analysis methods requires the involvement of a large community with diverse expertise and research goals. This section describes major components of the Linked Earth Platform that is currently supporting this community.

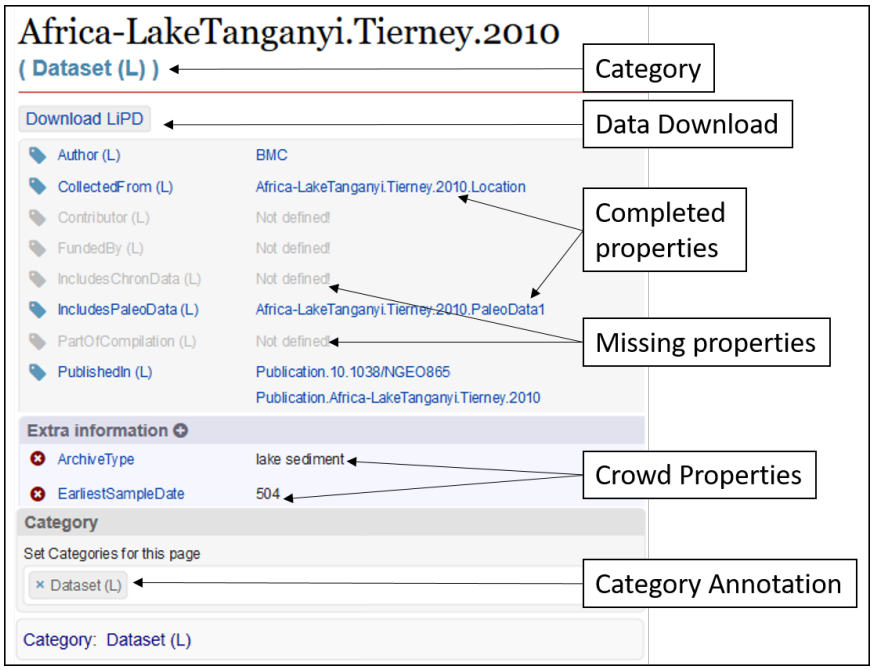

Figure 2: Overview of the metadata annotation interface, with core ontology terms marked with an "L". The properties under "Extra Information" are part of the crowd vocabulary.

4.1 Annotation Framework: The Linked Earth Platform

The Linked Earth Platform [21] implements the Annotation Framework as an extension of the Organic Data Science framework [9], which is built on MediaWiki [15] and Semantic MediaWiki [13]. There are several reasons for this. First, a wiki provides a collaborative environment where multiple users can edit pages, and where the history of edits is automatically tracked. Second, MediaWiki is easily extensible, allowing us to easily create special types of pages, generate dynamic user input forms, and create many other extensions. Third, because MediaWiki is well maintained and has a strong community, there are numerous plug-ins available. Finally, the Semantic MediaWiki API makes it easy to export content and interoperate with other systems.

Each dataset is a page in the Linked Earth Platform. "Dataset" is a special class, or category in wiki parlance. When a user creates a new page for a dataset, all the properties that apply are shown in a table where the user can fill their values. For each variable they indicate if it is observed or inferred, its value, uncertainty, and how it was measured. The order of the variables as columns in the data file is also specified.

Figure 2 shows the metadata annotation interface for a lake sediment dataset. The user has provided some of the values of the metadata properties, others have not been filled out yet. The core ontology properties are shown at the top, and the crowd vocabulary properties are shown near the bottom (under "Extra Information"). The user can also specify a new subcategory for this dataset, as shown at the bottom.

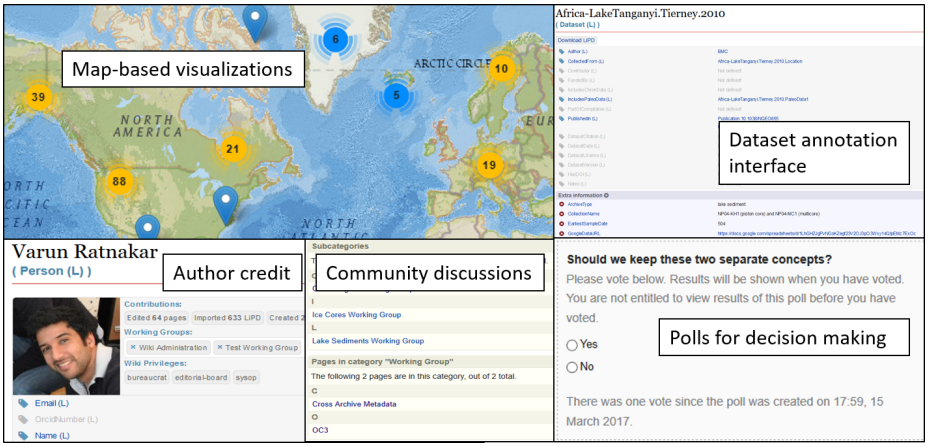

Figure 3: Overview of the main features of the Linked Earth Platform.

When annotating metadata, the system offers in a pull-down menu the possible completions of what the user is typing based on similar terms proposed by other users. This helps avoid proliferation of unnecessary terms and helps normalize the new terms created. If none represents what the user wants to specify, then a new property will be added. The property becomes part of the crowd vocabulary, and a new wiki page is created for it. The user, or perhaps others, can edit that page to add documentation. As a result, users build the crowd vocabulary while curating their own datasets.

Figure 3 highlights the main features of the Linked Earth Platform. The map-based visualizations show datasets already annotated with location metadata. Author pages show their contributions, which help track credit and create incentives. Other pages are devoted to foster community discussions and take polls. The annotation interface is designed to be intuitive, and provides detailed documentation with examples[1].

4.2 Initial Core Ontology

To ensure that most changes would be crowd extensions that would not cause major redesigns of the core ontology, the initial core ontology was carefully designed.

First, the ontology was developed using a traditional methodology for ontology engineering [23]. We started by collecting terms to be included by the ontology in collaboration with a select group of domain scientists. These terms where extracted from examples provided by the community[2], and from previous workshops where the community had discussed dataset annotation [4]. The ontology development process was also informed by previous efforts to represent basic paleoclimate metadata [14], and by prior community proposals to unify terminology in the paleo-climate domain [5].

[1] http://wiki.linked.earth/Best_Practices

[2] https://github.com/LinkedEarth/Ontology/tree/master/Example

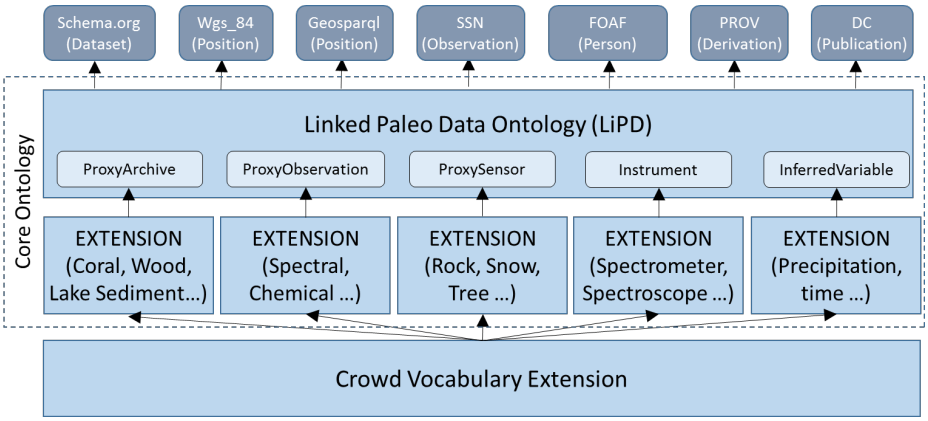

Figure 4: An overview of the core Linked Earth Ontology and its extensions.

We also took into account relevant standards and widely used models. We used several vocabularies[3]: Schema.org and Dublin Core Terms (DC) for representing the basic metadata of a dataset and its associated publications (e.g., title, description, authors, contributors, license, etc.), the wgs_84 and GeoSparql specifications for representing locations where samples are collected, the Semantic Sensor Network (SSN) to represent observation-related metadata, the FOAF vocabulary to represent basic information about contributors, and PROV-O to represent the derivation of models from raw datasets.

Figure 4 shows an overview of the ontology, which is layered and has a modular structure. The existing standards just mentioned provide an upper ontology for basic terms. We used the LiPD format, mentioned in Section 2, to develop the LiPD ontology[4] which contains the main terms useful to describe any paleoclimate dataset (e.g., data tables, variables, instrument used to measure them, calibration, uncertainty, etc.). A set of extensions of LiPD cover more specific aspects of the domain. The **Proxy Archive** extension defines the types of medium in which measurements are taken, such as marine sediments or coral. The **Proxy Observation** extension describes the types of observations (e.g., tree ring width, trace metal ratio, etc.) that can be measured. The **Proxy Sensor** extension describes the types of biological or non-biological components that react to environmental conditions and reflect the climate at the time. The **Instrument** extension enumerates the instruments used for taking measurements, such as a mass spectrometer. The **Inferred Variable** extension describes the types of climate variables that can be inferred from measurements or from other inferred variables (e.g. temperature). The crowd vocabulary builds on these extensions.

The core ontology and the crowd vocabulary share a common namespace for all the extensions (http://linked.earth/ontology/), in order to simplify querying as well as

[3] http://schema.org/, http://dublincore.org/documents/dcmi-terms/, https://www.w3.org/2003/01/geo/wgs84_pos, http://schemas.opengis.net/geosparql/1.0/geosparql_vocab_all.rdf, https://www.w3.org/2005/Incubator/ssn/ssnx/ssn#, https://www.w3.org/2005/Incubator/ssn/ssnx/ssn#, http://xmlns.com/foaf/spec/, http://www.w3.org/TR/prov-o/

[4] http://wiki.linked.earth/Linked_Paleo_Data

imports and exports of the ontology as a whole. Each extension has its own *base URI* e.g., http://linked.earth/ontology/instrument/), so they can be independently accessed

The Linked Earth ontology, as well as all the extensions, are accessible online[5] with content negotiation for HTML, RDF/XML and Turtle.

4.3 Community Organization and Support

The social aspects of the platform are equally important to the technical aspects. To organize the contributors' activities, we have created several mechanisms that are well documented and transparent to everyone. Our editorial processes were inspired by those of the Gene Ontology, Wikipedia, and our prior work on analyzing dozens of semantic wiki communities [10].

We have introduced four different user roles for crowd contributors. A *visitor* is any user who just wants to explore the content of the wiki. Visitors cannot change any of the contents of the wiki. By default, every user has a visitor role. A *basic editor* is a user with basic understanding on how to annotate a dataset with the Linked Earth Framework by creating a new page and adding new metadata, or by adding to a dataset created by someone else. Basic editors may also contribute to the textual documentation of existing terms from the crowd vocabulary, and may propose changes to the terms added by others. However, they cannot edit the semantics of the properties (e.g., domain and range). An *advanced editor* is a user with a more sophisticated understanding of the Linked Earth Framework and basic knowledge about ontologies. In addition to the basic editor privileges, advanced editors can add definitions for new properties, specify their semantics (e.g., domain and range, subclassing, etc.), suggest changes to existing properties proposed by others and reorganize the categories of the crowd vocabulary. Finally, an *editorial board member* is a user with extensive experience with the Linked Earth Platform, a deep understanding of the core ontology, and knowledge of the history of previous changes and issues discussed. Editorial board members are responsible for new versions of the core ontology, taking into account the possible ramifications of proposed changes by the crowd before incorporating them into the core ontology.

Users start as basic editors and then progress to advanced editors and in some cases may become an editorial board member. These user roles extend the default role functionality of MediaWiki, and are used extensively in collaborative content creation platforms such as Wikipedia [26].

In addition, we have set up *working groups* to organize activities by users with similar expertise. Led by advanced editors, working groups build use cases and examples and discuss problems raised by extensions to the crowd vocabulary. Agreements by a working group have more possibilities to convince the editorial board to accept their suggestions, as they represent the consensus of a set of experts in a particular domain instead of individual opinions. A working group is assigned a special page (of category Working Group) as a nexus for their activities. These include discussions and polls prompted by working group leaders with very specific questions

[5] http://linked.earth/ontology/

and choices for community voting. Polls are implemented through a MediaWiki plug-in[6], and advertised to the community through social media.

4.4 Ontology Revision and Update Framework

Once the editorial board agrees to create a new version of the core ontology, they start by generating a snapshot of the contents of the Linked Earth Platform. Then one person does the actual edits and updates using Protégé [24] for editing the core ontology, manually updating existing dataset descriptions, creating a new version of the wiki and using WIDOCO [7] and w3id.org[7] to document and do the content negotiation on the ontology. The new version of the core ontology is published online. Each ontology extension has a different version IRI, following the convention BaseURI/ExtensionName/VersionNumber (for example the Instrument ex, http://linked.earth/ontology/instrument/1.0.0), so they can be independently accessed. Finally, the editorial board updates the metadata annotations for datasets and the content of wiki pages through a semi-automated process.

4.5 Incentives

An important incentive for users is getting credit for all their contributions. Each of the contributions done by a user is tracked and shown in their profile page, summarizing how many pages the user created or edited, how many terms they have proposed, as well as the working groups where the user contributes. Details on the specific contributions are accessible on every user page. This is important for recognition of the work of different individuals, as well as to acknowledge contributions in publications.

Another key feature that we have incorporated to the platform is allowing scientists to upload entire metadata specifications along with datasets already created as LiPD files [14], rather than using the annotation interface. This is important because the LiPD format is supported by a research software ecosystem that help users manipulate and analyze paleoclimate data (e.g., GeoChronR[8], Pyleoclim[9]), and as a result, some scientists store their data in the LiPD format. We support batch import of LiPD files through a web interface[10], which assists users by ensuring that the data conforms to the LiPD descriptions, and then creates metadata and adds it to the Linked Earth Platform. The LiPD format is fully aligned with the core ontology,[11] so that downloading a LiPD file from any dataset in the wiki will contain the most recent updates done to its metadata.

---

[6] https://www.mediawiki.org/wiki/Extension:AJAXPoll
[7] http://w3id.org/
[8] https://github.com/nickmckay/GeoChronR
[9] https://pypi.python.org/pypi/pyleoclim
[10] http://lipd.net/create
[11] https://github.com/LinkedEarth/Ontology/blob/master/draftDiscussion/MappingLiPD-Le.xlsx

5 The Linked Earth Community Uptake

The Linked Earth Platform has been announced in several paleoclimate forums, receiving positive feedback from the community. The Linked Earth Platform is accessible online[12]. To date, the wiki has been populated with 692 datasets (mostly from the global PAGES 2k [18] collection[13]). We soon expect to increase this number with more than 150 additional records from a benthic oxygen isotope stack collection [1]. It is important to seed the platform with these datasets as a motivation for scientists to use the site, as well as a baseline for users to explore the capabilities of the system for querying and visualizing datasets.

Regarding user contribution, there are 150 registered users of the Linked Earth Platform (excluding the authors of this paper), with more than 50 contributors since the platform was first released. Users participate in one (or several) of the 12 available working groups, which tackle subjects in specialized topics such as how to unify cross-archive metadata and how to describe archive-specific metadata such as the storage conditions for an ice core or the coring technology used to obtain a sediment core. In order to facilitate decisions, working group members can respond to polls with specific questions. Previous votes for a particular decision are always recorded.

Figure 5 shows on the left the distribution of the main types of pages created to date. The total number of pages is more than 14,000. The right side of the figure shows the collaboration network of the main contributors of working groups. Each node represents a user, and each link between two users represents their collaboration on a given working group page (thicker lines signify several pages). The central nodes of this network are two of the authors of this paper (Jeg and Khider nodes), both editorial board members who help coordinate among the different working groups.

The vast majority of the edits to date have been annotations to the datasets using existing ontologies; a much smaller proportion has involved adding new terms. We have not deprecated terms so far, but they will be handled using current approaches: the term remains in the ontology and a warning to suggest what new related terms are recommended. We have only gone through two core ontology update cycles so far. The editors plan to do updates every six months, and will do them more often in periods of significant crowd vocabulary growth.

We are constantly working towards attracting more users to use the wiki framework. We have developed guidelines, tutorials and demos[14] on how to create content and add metadata. We have also showcased[15] how to query the datasets in the wiki through their metadata in order to analyze them with paleoclimate tools such as the Pyleoclim and the GeoChronR software mentioned earlier. We expect that these materials will help us increase the user base of the system.

[12] http://wiki.linked.earth/
[13] http://wiki.linked.earth/PAGES2k
[14] http://wiki.linked.earth/Best_Practices
[15] https://goo.gl/IGldxH

Page Distribution	
Datasets	659
ProxyAcrhive	207
ProxyObservation	76
ProxySensor	63
Instrument	45
InferredVariable	1207
MeasuredVariable	3348
Working Group	12
Location	659
Person	524
Publication	875

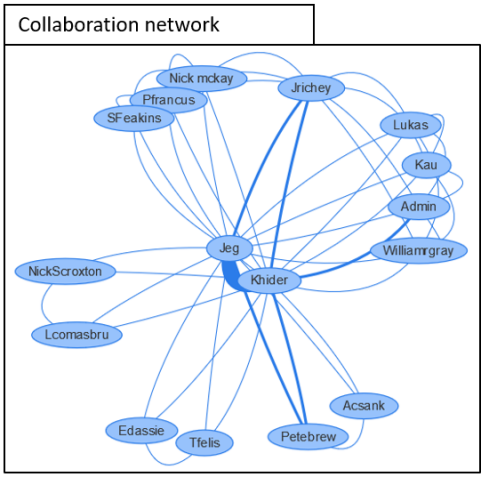

Figure 5: Use of the Linked Earth Platform. On the left we show the distribution of types of pages, while on the right we show the collaboration network of select users in working groups.

In a new project, we are adapting the Linked Earth Framework for a large international neuroimaging genomics collaboration to organize datasets and experiments from hundreds of different institutions worldwide.

6 Discussion

Linked Earth is still a young project. We have iterated once through the crowdsourced annotation and controlled revision cycle (current version is 1.2.0), creating new versions of the core ontology and the crowd vocabulary and updating the metadata of datasets accordingly. Working groups are already in the next cycle, discussing and voting on new terms to be added.

One could argue that having an editorial board to approve changes introduces delays in the updating process. However, we made sure that the initial core ontology was carefully developed with extensive community feedback to address major modeling issues, and we expect future changes will be relatively uncontroversial.

We continue to collect data about user contributions to the framework. We are particularly interested in the ability of the annotation interface to encourage the adoption of terms defined by others. We hypothesize the same effect reported in [6], where it was found that tag recommendation increases reuse across users, helps converge on a common vocabulary, and most importantly, promotes an increase in the quality of annotations.

An interesting request from the community was the ability to support publication embargoes, i.e., to keep a dataset private until the associated scientific publication is officially released. We have incorporated this feature so that selected datasets remain private until official release, though their metadata is always accessible to other users.

7 Related Work

Several collaborative frameworks have been proposed for knowledge engineering and ontology development, with a focus on either handling curation and enrichment of instances or refining the definitions of ontology concepts.

Several semantic wiki platforms support different forms of collaborative editing [3]. OntoWiki [2] is a collaborative wiki framework for editing and curating instances. The goal of OntoWiki is to help users editing the contents of a knowledge base using different views like maps and forms. OntoWiki is targeted toward the curation of instances, rather than describing an ontology.

Other wiki approaches focus on ontology development, with different features and in different domains. LexWiki [11] aims for a collaborative creation and categorization of taxonomies. Similarly, CSHARE [12] proposed to build an ontology to represent studies and experimentations. Moki [8] aims to capture more complex ontologies in the business processes domain. None of these approaches combines ontology editing with annotation of instances.

Collaborative ontology editors target users with expertise on the Semantic Web. As an example, Collaborative Protégé [24] is a full ontology editor, which includes features for enabling discussion, comments and annotations of ontologies in a distributed manner. These features also appear in Web Protégé [25], a lightweight ontology editor for the Web. Both versions can be used for defining instances of the ontology classes (e.g., through forms). SOBOLEO [27] is a collaborative vocabulary editor designed for organizing lightweight SKOS taxonomies that can annotate external web pages with the concepts in the taxonomy. PoolParty [22] is a wiki editor designed to edit and augment SKOS thesauri combined with the ability to process documents and look in external datasets for new concepts to add. However, none of these editors support the fast-paced cycle for ontology extension and immediate use for annotation in our approach.

The Diligent methodology [20] allows users to create individual extensions of a common core ontology, and these separate extensions are then merged by ontology engineers. In our work, all users work with the same extension to the core ontology, working collaboratively to create new terms.

8 Conclusions

We have presented a novel socio-technical approach for ontology development and data annotation based on controlled crowdsourcing. Key aspects of this approach are: 1) a crowdsourcing annotation process that allows users to add new terms as they use a standard ontology to do data annotation; 2) an editorial revision process that incorporates new terms into the next version of the ontology; and 3) a framework for updating the annotations according to that new version. We have implemented this approach in the Linked Earth Platform for the paleoclimate community. Seeded with an initial core ontology, the platform is being used to extend that ontology as needed by scientists as they annotate their datasets to create a crowd vocabulary. Although it is

still early on in the project, the community is actively engaged in proposing terms and revising the core ontology. Future work includes facilitating ontology convergence, formalizing types of ontology changes to facilitate automation of updates to the repository, and improving update documentation and tracking.

Acknowledgements

We gratefully acknowledge funding from the US National Science Foundation under EarthCube grant ICER-1541029 and under grant IIS-1344272. We would like to thank the paleoclimate scientists who are participating in this community effort. We also thank Chris Duffy, Paul Hanson, Jie Ji, Tejal Patted, and Neha Suvarna for their contributions to the project.

References

1. Ahn, S., Khider, D., Lisiecki, L., and C. E. Lawrence. A probabilistic Pliocene-Pleistocene stack of benthic δ18O using a profile hidden Markov model. *Dynamics and Statistics of Climate Systems*, 2017. doi: 10.1093/climsys/dzx002
2. Auer, S., Dietzold, S., and T. Riechert. OntoWiki – a tool for social, semantic collaboration. *Proceedings of the International Semantic Web Conference*, 2006.
3. Bry, F., Schaffert, S., Vrandečić D., and K. Weiand. "Semantic Wikis: Approaches, Applications, and Perspectives." Lecture Notes in Computer Science, *Reasoning Web. Semantic Technologies for Advanced Query Answering*, Volume 7487, 2012.
4. Emile-Geay, J. & McKay, N. P. Paleoclimate data standards. *Pages Magazine* 24:1, 2016. doi:10.22498/pages.24.1.47
5. Evans, M. N., Tolwinski-Ward, S. E., Thompson, D. M., & Anchukaitis, K. J. Applications of proxy system modeling in high resolution paleoclimatology. *Quaternary Science Reviews*, 76, 16-28, 2013. doi:10.1016/j.quascirev.2013.05.024
6. Font, F., Serrà, J. and X. Serra. "Analysis of the Impact of a Tag Recommendation System in a Real-World Folksonomy". *ACM Transactions on Intelligent Systems and Technology*, (7)1, 2016.
7. Garijo, D. WIDOCO: A Wizard for Documenting Ontologies. *Proceedings of the Sixteenth International Semantic Web Conference (ISWC)*, 2017.
8. Ghidini, C., Kump, B., Lindstaedt, S., Mabhub, N., Pammer, V., Rospocher, M. and Serafini, L. MoKi: The Enterprise Modelling Wiki. *Proceedings of the 6th Annual European Semantic Web Conference (ESWC)*, 2009.
9. Gil, Y., Michel, F., Ratnakar, V., and Hauder, M. Organic Data Science: A Task-Centered Interface to On-Line Collaboration in Science. *Proceedings of the ACM International Conference on Intelligent User Interfaces (IUI)*, 2015.
10. Gil, Y.; and Ratnakar, V. Knowledge Capture in the Wild: A Perspective from Semantic Wiki Communities. *Proceedings of the Seventh ACM International Conference on Knowledge Capture (K-CAP)*, 2013.
11. Jiang, G. and Solbrig, H. R. Lex Wiki framework and use cases. *First meeting of Semantic Media Wiki users*. November 22-23, 2008. Boston, MA, USA.

12. Jiang, G., Solbrig, H. R., Iberson-Hurst, D., Kush, R. D., & Chute, C. G. (2010). A Collaborative Framework for Representation and Harmonization of Clinical Study Data Elements Using Semantic MediaWiki. *Summit on Translational Bioinformatics*, *2010*, 11–15.
13. Krötzsch, M., Vrandečić, D.: Semantic MediaWiki. *Foundations for the Web of Information and Services* – A Review of 20 Years of Semantic Web Research, pp. 311–326. Springer, 2011.
14. McKay, N. P. and Emile-Geay, J. Technical note: The Linked Paleo Data framework – a common tongue for paleoclimatology. Climate of the Past, 12:4, pp 1093-1100, 2016. doi: 10.5194/cp-12-1093-2016
15. MediaWiki, The Free Wiki Engine. 2017. https://www.mediawiki.org.
16. Nielsen M. *Reinventing Discovery*. Princeton University Press, 2011.
17. PAGES 2k Consortium. Continental-scale temperature variability during the past two millennia, *Nature Geoscience*, 6(5), 339–346, 2013. doi:10.1038/ngeo1797.
18. PAGES 2k Consortium. A global multiproxy database for temperature reconstructions of the Common Era, *Scientific Data*, 4, 170,088 EP, 2017. doi:10.1038/sdata.2017.88.
19. Pangaea Interoperability and Services. 2017. Available from https://pangaea.de/about/services.php
20. Pinto, H.S., Tempich, C., Staab, S., Sure, Y.: Distributed Engineering of Ontologies (DILIGENT). In: *Semantic Web and Peer-to-Peer*, Springer, Heidelberg, 2005.
21. The Linked Earth Wiki. V. Ratnakar. [Dataset]. *Zenodo*. http://doi.org/10.5281/zenodo.579646
22. Schandl, T., and Blumauer. A. PoolParty: SKOS thesaurus management utilizing linked data. *Proceedings of the 7th international conference on The Semantic Web: research and Applications (ESWC)*, 2010.
23. M.-C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, and A. Gangemi. *Ontology Engineering in a Networked World*. Springer, Berlin, 2012
24. Tudorache T., Noy N.F., Tu S., Musen M.A. (2008) Supporting Collaborative Ontology Development in Protégé. *Proceedings of the International Semantic Web Conference (ISWC)*, 2008.
25. Tudorache,T., Nyulas, C., Noy, N.F., and Musen, M.A. WebProtégé: A collaborative ontology editor and knowledge acquisition tool for the Web. *Semantic Web* 4(1), 89-99, 2013.
26. Wikipedia, The Free Encyclopedia. 2017. https://en.wikipedia.org/
27. Zacharias, V. and Braun, S. SOBOLEO - Social Bookmarking and Lightweight Ontology Engineering. *Workshop on Social and Collaborative Construction of Structured Knowledge, 16th International World Wide Web Conference, 2007.*