Organic Data Publishing: A Novel Approach to Scientific Data Sharing

Yolanda Gil and Varun Ratnakar
Information Sciences Institute and Department of Computer Science
University of Southern California
`gil@isi.edu`

Paul C. Hanson
Center for Limnology
University of Wisconsin-Madison
`pchanson@wisc.edu`

Abstract. Many scientists do not share their data due to the cost and lack of incentives of traditional approaches to data sharing. We present a new approach to data sharing that takes into account the cultural practices of science and offers a semantic framework that 1) links dataset contributions directly to science questions, 2) reduces the burden of data sharing by enabling any scientist to contribute metadata, and 3) tracks and exposes credit for all contributors. To illustrate our approach, we describe an initial prototype that is built as an extension of a semantic wiki, can import Linked Data, and can publish as Linked Data any new content created by users.

Keywords: Scientific data sharing, provenance, semantic wiki, Linked Data.

1   Introduction

Although scientists in many disciplines share data through catalogs so that others can harvest those data for analysis and publications (e.g., in astronomy, physics, etc), this paradigm has not worked well in ecology. Ecology is a field science, where many scientists have their own data collection instruments and often curate datasets themselves for a particular location for many years. Vast amounts of data are sitting on local systems of many thousands of scientists, often called "dark data" [Heidorn 2008]. These datasets are often very specific to a locality or phenomenon, but they are developed by the vast majority of scientists, known as "the long tail of science." Some report that less than 1% of data in ecology are available once they are analyzed and results are published [Reichman et al 2011]. Although scientists would like to share data, they often do not do so for four fundamental reasons [Science 2011]:

1.   the paradigm makes data providers second class citizens, and for some ecologists, data are a primary asset

2. data in ecology are complex, highly distributed and typically obtained to answer local questions, and posting those data in ways that make them discoverable and accessible requires a lot of work
3. the probability of posted data being discovered independent of the science social network is low, reducing greatly the motivation to post the data
4. almost all data sharing today begins with scientific collaboration, and traditional data sharing approaches are not linked to collaborative activities

Many current projects in geosciences depend on having broad access to data from the long tail of science. Many observatory networks and initiatives such as EarthCube[1] envision the geoscience community coming together to ecosystem, regional, continental, and global-scale problems. For example, to understand the carbon cycle in water involves integrating data and analyses by scientists studying river, lake, ocean, and coastal ecosystems. Critical research in ecology and geosciences can only be addressed through the integration of data and models from thousands of scientists spanning many disciplines (ocean, earth, and atmospheric sciences).

These projects need the data to be shared, but furthermore the data needs to support ad-hoc data sharing and collaborations. The data needs to be openly available and well annotated with metadata so it can be aggregated and integrated. We need approaches that part with the artificial walls created by traditional discipline-specific data catalogs and infrastructure projects.

We are investigating *organic data sharing* as a novel approach to data publishing that is open to all scientists to contribute in many forms, requires minimal effort from contributors, collects and exposes credit for all contributions, and has emergent organization. Our work builds on three interrelated techniques: semantic web standards, linked web of data principles, and popular web paradigms for interfaces such as semantic wikis to annotate and aggregate data.

This paper describes our initial work towards this vision. We begin with an overview of the approach, followed by a walkthrough of a prototype that we have developed to illustrate it. We also present a visionary scenario that shows how this approach would open science to a broader set of contributors.

2 Organic Data Sharing

Organic data sharing builds on three interrelated techniques:

1. *Semantic web standards* for defining semantic metadata in an extensible way over web standards, including the use of RDF to define data types and properties, which allow users either to reuse properties already defined in the system or to easily add and use new properties.
2. *Linked data principles* to expose datasets and their semantic metadata in an open form on the Web. Traditional data repositories will upload data to a central or distributed database, akin to a vault where the data is kept. In contrast,

---

[1] http://www.nsf.gov/geo/earthcube/

linked data principles encourage all data and metadata to be web objects that can be openly accessed by third-party web applications. There are vast and rapidly growing amounts of linked data published in this format. They already include large amounts of datasets relevant to ecology, such as geospatial data (Geonames, OpenStreetMap), life sciences data (Gene Ontology, PDB), and academic publications (PubMed, ACM), and Wikipedia info boxes (DBPedia).

3. *Semantic wikis* as popular web paradigms for interfaces and access to facilitate the creation of simple tools of broad applicability to browse, visualize, annotate, and integrate data. Semantic wikis augment traditional wikis so that the hyperlinks between topic pages are annotated with a semantic relationship. The contributors themselves can create the emergent structure of the content by adding new properties in an as-needed basis.

Our approach is to design an environment that supports scientists to carry out the following activities:

- any scientist can define collaborative tasks by stating questions that require participation from the broader community
- any scientist can contribute to those tasks, decompose them into subtasks if appropriate, and request particular kinds of datasets
- scientists can contribute datasets that they own simply by adding a pointer to their datasets which will continue to reside in their local systems and under their control
- any scientist can add metadata to any datasets, defining new metadata properties or adopting properties that others have used (or from common ontologies)
- any scientist can change the metadata specified for any dataset in order to adopt the same properties that other similar datasets use, facilitating aggregation of data
- any scientist can use any dataset, and must post the results of their analyses with appropriate links to the original datasets that they used

The system will support organic data sharing by:
- assigning credit to each individual scientist by tracking, aggregating, and exposing all their contributions of any nature
- pointing scientists towards tasks that could use their contributions by analyzing the semantic properties available
- allowing users to import content that may be available as linked data
- publishing as linked data any content created by users

3   An Illustration of Organic Data Sharing

This section illustrates organic data publishing through an initial prototype that extends a semantic wiki framework. Semantic MediaWiki builds on the popular MediaWiki software, and extends them to allow users to express semantic relations[2]. We

---
[2] http://semantic-mediawiki.org/

describe how the user interacts with the system in order to illustrate the capabilities of the system.

Figure 1 illustrates the variety of entities that can be linked to one another through structured properties. In the figure, one window shows a wiki page for a dataset, including semantic metadata properties that describe the collection instrument, location, and time as well as the investigator who contributed it. That location happens to be a lake, which is described in its own wiki page showed in a separate window in the image, with its own geospatial and other semantic metadata properties. A third window shows the wiki page for the investigator showing other contributed datasets and other information that might provide context for the data. Anyone can edit the wiki, add any metadata properties, extend metadata vocabulary, etc. All the information collected through the site is published as Linked Data.

All the contributors to each topic page are acknowledged, and there is a clear link to the scientist that contributes each original dataset.

The system enables contributors to easily define structured semantic properties to describe the contents of the wiki, and uses RDF as the semantic representation standard. Each wiki page describes an object of interest (eg, a dataset, a project) and has a section of "Structured Properties", where contributors can specify properties and values of the topic of the page. Any contributor can define new properties on the fly. Any contributor can change an existing property to align it with one that is used elsewhere, effectively normalizing the use of the property across pages and therefore across objects. This results in an organic normalization of metadata properties for datasets, which would typically result when datasets need to be aggregated for some science purpose. Figure 2 shows an example of how the system creates content of wiki pages dynamically through queries, in this case a query to show three properties of lakes. Users browsing the site are immediately exposed to missing information and can choose to contribute it. When the missing information stands in the way of progress, they can be more motivated to add it.

The framework has pre-defined categories of pages, each with their own with pre-defined areas. We have defined so far five special categories: Question, Answer, Data, Workflow, and ExecutedWorkflow.

Figure 3 illustrates the special page category of Question. These are pages that reflect a task or subtask. They have sub-questions that point to pages of category Question as well. These subquestions may lead to request a dataset, as is the case in the example shown in the figure. Some workflows may be designed and later executed once the desired datasets are collected. When the question is answered, users can create a page of another category, Answer, that would summarize all the findings and perhaps include pointers to a publication. As any other page, question pages can have structured properties, and each is credited to its author.

Figure 4 illustrates the special page category of Data. These pages represent a dataset, which can have as always structured properties. Some properties, as is the case here, may be imported by the system from assertions available as Linked Data. Some sections of the page are created dynamically through queries, for example to show what workflows use the dataset as input (shown in orange in the figure).
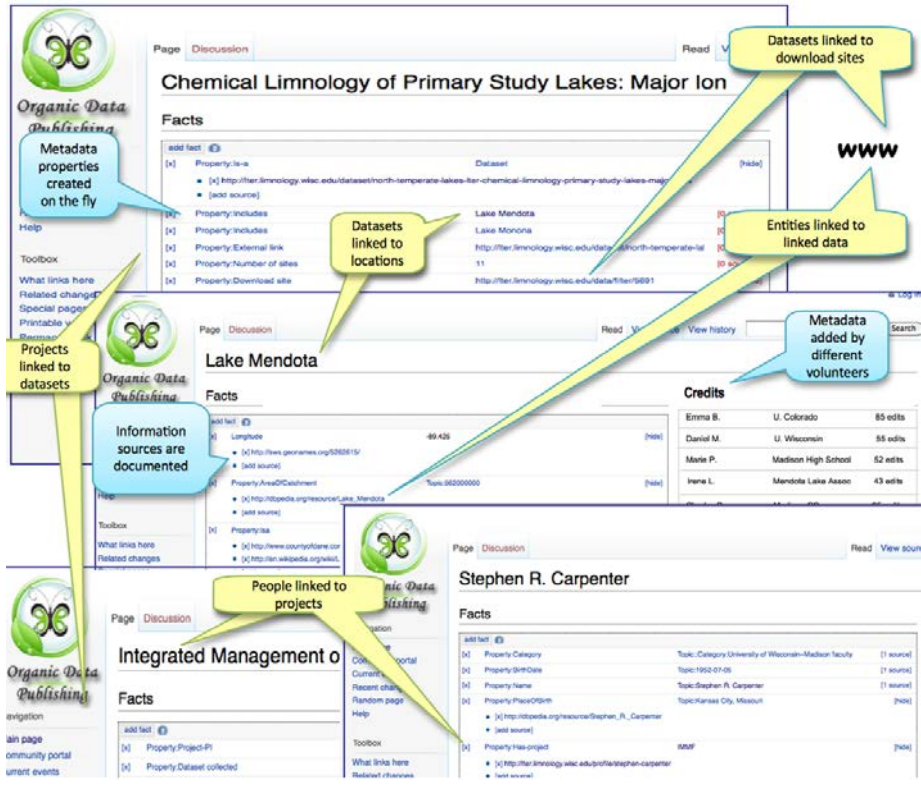
Figure 1. Overview of an organic data publishing site, implemented as a Semantic Wiki. Datasets, scientists, projects, and locations each have their own properties and are all interlinked. Datasets are linked to their download locations, and any object can be an object in Linked Data. Contributors are acknowledged for providing metadata for datasets, as well as properties and links for any objects.

	Area of Catchment	Latitude	Longitude
Lake Casitas	100,000,000	34.392	-119.335
Lake Mendota	562,000,000	43.107	-89.425
Lake Monona			
Lake Wingra		43.053	-89.422

Figure 2. Enticing users to contribute by exposing unknowns. With a semantic wiki, a query can be created to generate dynamic content on any wiki page. Shown here is the dynamic content created in answer to a query about lakes. Note that the entries in the table that are empty show users where they can contribute.

Global distribution of carbon in lakes

Answers to this Question

Add

- [x] Carbon budget for selected lakes

Sub Questions

Add

- [x] Calculate carbon budget for selected lakes
    - Calculate carbon budget for Lake Mendota
    - Calculate carbon budget for Lake Winga

- [x] Calculate CO2 levels for the air around the lake

Some References

Distribution of Labile Dissolved Organic Carbon in Lake Michigan

http://www.jstor.org/pss/2837545

*Biossay-measured, labile dissolved organic carbon (LDOC) concentrations were compared
between April and October 1986. In five of seven experiments, the LDOC concentration wa
of the total DOC pool in the near-bottom water in late May and 13.8% in the near-surface w
was highest during early stratification; concentration in surface water varied less but was hi
allochthonous source of labile organic C may be important.*

Structured Properties

Add

[x]	Gas flux publications	http://www.jstor.org/ps...	(By Hanson)
[x]	Level of difficulty	High	(By Gil)
[x]	Number of expected publications	20	(By Hanson)

Credits

Users who have contributed to this Question, its SubQuestions and Answers:
- Hanson (35 edits)
- Gil (4 edits)

▷ See details

Category: Question

Figure 3. A question (or task) can be decomposed, each subtask addressed in its own page. Text can be added as background documentation, when appropriate structured properties are associated with the question. Credits are shown prominently.

Figure 5 shows an example of a page with a special category of Workflow. In this case, the workflow was created using a separate workflow system, Wings[3], that publishes workflows as Linked Data using OPMW[4], an extension of the Open Provenance Model [Garijo and Gil 2011]. The system imports the OPMW assertions and shows the workflow in a wiki form. Again, anyone can add structured properties or documentation to this page.

---

[3] See http://www.wings-workflows.org
[4] See http://www.opmw.org

CDEC WEATHER 2010 03 02

Data

- **DOWNLOAD**
- **Data Types**
  - Daily Sensor Data
- **Used as Input in the following Workflows:**
  - AF NTM Execution 2 March 2012 to 8 March 2012
  - AF EDM Execution 2 March 2012 to 8 March 2012
  - AF EM Execution 2 March 2012 to 8 March 2012
  - AF NTM Execution 2 March 2012 to 31 March 2012
  - AF EDM Execution 2 March 2012 to 31 March 2012
  - AF EM Execution 2 March 2012 to 31 March 2012

Structured Properties

Add

[x]	Barpress	760	(By Admin)
[x]	Depth	1.0214570760727	(By Admin)
[x]	Flow	1550.6185302734	(By Admin)
[x]	ForSite	SMN	(By Admin)
[x]	HasSize	8316	(By Admin)
[x]	SiteLatitude	37.347213745117	(By Admin)
[x]	SiteLongitude	-120.97618103027	(By Admin)
[x]	Slope	0.0000999999997473788	(By Admin)
[x]	Velocity	0.65311223268509	(By Admin)

Credits

Users who have contributed to this Page:

- Admin (19 Edits)

Category: Data

Figure 4. A dataset can be imported by the system (Admin) together with its properties. The semantic wiki can dynamically generate content for pages, such as the workflows that use this particular dataset as input.

Figure 6 illustrates the ExecutedWorkflow category, showing also a page generated for one of the workflow data products. The structured properties in this case were imported by the system from assertions in Linked Data that were generated originally in Wings based on the properties of the datasets that were input to the workflow. A workflow execution can be linked to the appropriate question page.

Figure 5. A workflow, published by another system (Wings) as Linked Data, is imported so that it can be further annotated with properties and linked to other objects, such as questions, sub-tasks, datasets, or researchers. All these annotations are published in turn as Linked Data.

The framework incorporates the following major extensions to the semantic wiki:

- Contributions are driven towards answering global science questions is a great incentive for participation of scientists. Answering these questions will be the overarching goal, which will require contributors to do a variety of tasks such as decomposing the high level questions into smaller tasks, sharing datasets, describing data characteristics, preparing them, running models, etc.

Figure 6. A workflow execution, published by another system (Wings) as Linked Data, is imported and includes links to new datasets (the number of each type is indicated in parenthesis) generated by the workflow. Each dataset has structured properties that the workflow system propagated from the workflow's input datasets, and those properties are imported as well. The workflow execution can then be linked to the appropriate Question page.

- Workflow technologies and provenance standards are embedded in the framework to enable scientists to describe analytic processes that will document new data products in terms of how they were obtained from raw data. Worfklows are imported into the framework from Linked Data, where they are published by the workflow system that created them. Workflows and

their results could also be added manually by users, for example if the steps are run by hand or through scripts.

- Credit is given explicitly in every page and for every contribution. Credit is aggregated per question and per user. Wikis provide a natural infrastructure to track contributions, but they are typically hidden in the history tab of each wiki page. The contributor of a dataset can see what question it is contributing to and in what form (through the workflows that are using it).

We continue to extend this prototype to exemplify the approach of organic data sharing. We are working with the EarthCube community to identify additional requirements from scientists. More research is needed regarding contributor credits and data citations. We plan to explore different incentive and reward mechanisms that will suit the contributor's communities of practice. Another aspect we plan to investigate is the viability of emerging semantics as the contributors normalize the attributes and properties they use. We will analyze the drivers for convergence on semantic properties, the practical reuse of community ontologies such as SWEET, and their effect on productivity and data reuse.

4   Discussion

Quantitative data can be collected by instrumenting the system. We can use standard wiki data collection metrics used in studies of wiki user behaviors and content growth (e.g., the number of edits per user). We can also metrics particular to semantic wikis (e.g., the number of structured properties defined).

In addition to these more traditional wiki-style evaluations, we will be developing science-relevant metrics such as the number of datasets collected and the number of datasets aggregated through normalization of metadata properties. Another a novel aspect involved in the evaluation of the system revolves around task decomposition, task contributions, and task accomplishment that have not been addressed in prior work on contributor involvement.

We will need to explore alternative designs for the task-centered aspects of the approach. Recent work on social creation of to-do lists offers an alternative approach to creating and organizing subtasks [Kamar et al 2012]. Other successful examples for enticing contributors to contribute to joint tasks have used common collaborative web software [Rocca et al 2012]. Formative evaluations to compare these approaches could be carried out to determine what works best for organic data sharing.

We have identified four important dimensions of evaluation that are of interest: participation, collaboration, convergence, and achievement of the community.

Participation metrics can be used that are indicative of the involvement of users from the community. We can create an estimate of the size of the community as the total number of unique users who ever visit the site. The system can then collect the total number of users who edit pages and contribute content to the site, the total number of datasets contributed, and the total number of edits both collectively and per user. Additional, participation metrics can be collected regarding the structured prop-

erties defined in the semantic wiki, including the number of semantic properties added by user and the number of semantic properties defined for each type of dataset.

Collaboration metrics can indicate how users overlap in their activities as they collaborate on specific topic pages in the wiki. Data can be collected regarding number of users who edit the same topic page, the number of links across topic pages, and the number of users that contribute to a given stated task or subtask.

Convergence metrics will expose how the community normalizes structured properties as the metadata is added for the diverse datasets. These metrics can include the number of common properties across datasets used in a given task or workflow, amount of unique users that adopt each property, the number of deprecated semantic properties that are replaced by new (more broadly used) ones, and the evolution of semantic properties over time. In addition, the amount of queries defined in wiki pages to create dynamic content based on semantic properties would be an indicator that the content is being aggregated across separate pages and contributors.

Achievement measures the progress and accomplishments of task-oriented contributions. The system can collect metrics regarding the amount of tasks and subtasks created, the amount of data collection and workflow pages created associated with tasks, the amount of user activity associated with each task and with wiki pages over time, and the amount of subtasks with answers as indicators of accomplishment.

We plan to extend the system to take on a more proactive role in soliciting contributions. The system could do meta-analyses on the content at any given point in time, determine what is needed, and prompt users accordingly. For example, it could determine what tasks have not advanced for some time, propose decomposing them into smaller subtasks that define contributions more specifically, and identify who could be approached to make a specific needed contribution based on their past history.

Central to our approach is the tracking and exposure of credit to individual contributors on a topic-by-topic as well as an individual basis. It is important for the system to track contributions of any size and nature, ranging from contributions that require significant effort (e.g., the contribution of a dataset that took months to collect), to very small effort (e.g., the renaming of a property of a dataset to standardize names across datasets), and any effort in between (e.g., the addition of a metadata property to a dataset that required analyzing the data to decide on the property value). Another important aspect of the system is to reflect the credit for user contributions whenever content is presented, whether it is overall user credit in a user page, or ranked credits to all users for a given topic page. Ranking contributors in scoreboards appears to be a great incentive in social computing systems, and we will explore this.

For owners of datasets (the dark data from the long tail), the explicit links from the data to the scientific problems it is used for will address the concern of the recognition of their contributions to problems. Another issue that this will address is that they will be able to inspect that their data is used for appropriate goals and with appropriate transformations to fit the models used in the analyses. A benefit for them will also be that their future data collection efforts will put them in a position of being able to re-run the analyses with the new data. Currently, they typically lack the knowledge about how to run models as well as access to their codes. These issues will be addressed by the availability of the analyses in the system.

In the end, the credit tracked and acknowledged in our system must be recognized in the traditional forms of credit in science as scientific publications. The credit tracking in our approach will have to be combined with social rules that set expectations about how contributors are acknowledged in any resulting publications. An approach taken in Polymath is that the author is named as "Polymath" and a pointer to the web site is provided where all contributors are acknowledged in detail together with the nature of their contributions. We will explore together with the scientists in the community what would be appropriate acknowledgements in publications.

5 Conclusions

We presented organic data sharing as a novel approach to collect dark data from the long tail of science in a form that can be enticing to scientists and including metadata annotations that make the data most usable. There are many potential benefits of the proposed approach: 1) the publication of data and metadata is virtually instantaneous, so is its access; 2) each scientist is personally responsible and in charge of the publication of their data; 3) scientists, students, citizens, and policy makers can all be contributors; 4) data descriptions can be created in an ad-hoc manner, and normalized and integrated in an as needed basis; 5) everyone else benefits when someone invests in describing, normalizing, or aggregating data; and 6) the immediate benefits to each scientist should be enough to make data publishing and metadata creation become a pleasant habit rather than a chore.

There is already a success story of scientific sharing that has these properties. The Web has all these properties: instantaneous, personal, participatory, self-organizing, empowering, and addictive. We are building on web infrastructure to foster the creation of a web of data for environmental science.

References

1. Garijo, D., and Gil, Y. "A New Approach for Publishing Workflows: Abstractions, Standards, and Linked Data". In Proc. WORKS'11, Seatle, WA, 2011.
2. Heidorn, P.B. "Shedding Light on the Dark Data in the Long Tail of Science." Library Trends, Vol. 57, No. 2, Fall 2008.
3. Kamar, E., Hacker, S. and E. Horvitz. "Combining Human and Machine Intelligence in Large-scale Crowdsourcing," AAMAS 2012, Valencia, Spain, June 2012.
4. Reichman, O.J., Jones, M.B., and M.P. Schildhauer. "Challenges and Opportunities of Open Data in Ecology." Science, Vol. 331 no. 6018 pp. 703-705, February 2011, DOI: 10.1126/science.331.6018.692.
5. Rocca RA, Magoon G, Reynolds DF, Krahn T, Tilroe VO, et al. "Discovery of Western European R1b1a2 Y Chromosome Variants in 1000 Genomes Project Data: An Online Community Approach." PLoS ONE 7(7), 2012.
6. Science, 2011, Special Issue on Challenges and Opportunities. Vol. 331 no. 6018 pp. 692-693, February 2011, DOI: 10.1126/science.331.6018.692.