WEB SYNDICATION APPROACHES FOR SHARING PRIMARY DATA IN "SMALL SCIENCE" DOMAINS

***Eric C. Kansa**[1*] **and Ahrash Bissell**[2]*

*[*1]School of Information, UC Berkeley, 102 South Hall Road, Berkeley, CA 94720-4600 (USA)
Email: ekansa@ischool.berkeley.edu
[2]The William and Flora Hewlett Foundation
Email: ahrash.bissell@gmail.com*

ABSTRACT

In some areas of science, sophisticated web services and semantics underlie "cyberinfrastructure". However, in "small science" domains, especially in field sciences such as archaeology, conservation, and public health, datasets often resist standardization. Publishing data in the small sciences should embrace this diversity rather than attempt to corral research into "universal" (domain) standards. A growing ecosystem of increasingly powerful Web syndication based approaches for sharing data on the public Web can offer a viable approach. Atom Feed based services can be used with scientific collections to identify and create linkages across different datasets, even across disciplinary boundaries without shared domain standards.

Keywords: eScience, Cyberinfrastucture, Atom Syndication Format, Linked Data, Semantic Web, Small Science, Field Science, Open Data, Open Access, Data Sharing

1 INTRODUCTION

In 1898, Hermon Bumpus published a landmark study on the evolutionary process of stabilizing selection by investigating mortality among house sparrows. Unlike most of his contemporaries, he comprehensively published his primary observations along with his theoretical interpretations. His raw data have been tremendously valuable to later researchers, and inspired publication of many influential, peer-reviewed papers. Measuring the value of raw data by the number of publications it spawns, this set of raw data was at least 10 times more valuable than it would have been without dissemination. Such reuse can increase dramatically if the raw data are available on general public networks such as the Internet (Kansa et al., 2005). The Bumpus data have even more value when considering their use for student instruction and exploration of "real world" evidence. As practices of data discovery and reuse change, researchers will increasingly see the benefit of providing greater and more transparent access to primary field datasets and specialist analyses (see Hajjem et al., 2005; Kintigh, 2006; Onsrud & Campbell, 2007; Snow et al., 2006; Piwowar et al., 2007). Such datasets will be judged by their significance and by their ability to support future scholarship (Richards, 2004).

Impediments toward more open and comprehensive dissemination of data include a variety of professional, conceptual, and technological challenges common to many "small sciences" (Onsrud & Campbell, 2007). Small-science domains include field sciences such as archaeology, ecology, and conservation biology, as well as some areas of public health, education, and psychology. Small science typically works in decentralized institutions with case-specific research questions, often using customized methods and recording systems and individually maintained data resources (Borgman et al., 2007). In an example typical of many small science domains, archaeologists generally adhere to few specific methodological or recording standards, and often make customized databases to suit their individual research agendas focusing on local or regional-scale research (see also Borgman et al., 2006, 2007; Baru, 2007). For example, the particular nature of archaeology, a discipline straddling the humanities, social sciences, and natural sciences, necessitates a diversity of documentation needs and methods. Data, evidence, interpretations, and syntheses have different roles across this widely varying community (for discussion of social-science material, see Paterson, 2003). For example, one excavates and interprets a Paleolithic cave site differently from a Roman urban site. Also, practical and budgetary factors external to scientific aims are important in shaping documentation strategies. Much archaeological research takes place as part of heritage or cultural resource management. Excavation sampling strategies and laboratory analyses may all be shaped by construction timelines and imperatives, permitting requirements, property owners, and community interest groups. As a consequence, archaeological excavation results, specialist analyses, and museum collections are highly variable (Kintigh, 2006).

2 ACHIEVING CRITICAL MASS: THE QUALITY OF QUANTITY

The current landscape sees small-science data collections fragmented among various content silos. This fragmentation hampers demonstration of clear data sharing benefits. Most online data-sharing initiatives have primarily focused on developing "destination web sites." Such sites implicitly put people in the position of being a passive audience, accessing content through a predefined user interface offered by the site.

Such approaches typically leave content in splendid, and sometimes beautifully designed, isolation. Usually, little thought is given to making collections open for interaction through alternative interfaces, or for making content open for use in alternative contexts. Potentially, other data published on the Web can allow searchers to augment and further contextualize content published in a given collection. However, few small-science data publishers offer alternative interfaces, especially Web-services and APIs ("application program interfaces") that would allow content to escape a given silo so that it could be used and compared to data published by other sources.

While the "destination web site" approach does offer access to research content, it does not fundamentally transform how the content can be used. One of the great potentials for data sharing is enabling wholly new research programs unprecedented in scope and analytic rigor. But this goal cannot be realized simply through access to data. Data need to be portable and open for aggregation and comparative analysis with other datasets, some of which may come from outside the disciplinary boundaries of a particular field. Individual interests can be quite specific and even esoteric. Typically, no one single collection grows large enough to obtain a "critical mass" required to sustain wide interest. More often, bodies of relevant content will be published by several different sources. If one cannot effectively work across these different sources, the network effects expected from data sharing will never be realized.

Thus, a key issue for the community to finally enjoy the research benefits of data sharing is to enable the data to reach the researchers, rather than expect the researchers to find the data. In other words, you may not be able to provide the best and most innovative way of presenting and using your collection. But if your data are open and portable, someone else may be able to provide additional content and context that can unlock hidden value. Data portability is thus a key requirement for building a critical mass of useful content necessary for enabling transformative research.

3 THE ROLE OF STANDARDS

Organizations and teams that publish data include museums and research organizations, field stations, government archives, or small university-based groups. All of these groups use their own data models, recording systems, taxonomies, etc. Although they may publish data to the Web, even well financed organizations may not see a compelling need to invest money and effort to express structured data according to a common, domain-specific standard. We should expect significant decentralized publication and dissemination of small-science data and a diversity of structured data standards well into the future.

The diversity of scientifically relevant data collections is more than a function of different funding levels and technical capacities. For example, in archaeology the wide scope of interests and disciplinary inputs is an even greater driver for data diversity (see related Fry & Talja, 2007). Projects like Nomisma.org and Pleiades have expertly collected and maintained numismatic and historical geography collections, outside the traditional scope of "field archaeology." Beyond these different disciplinary interests and worldviews, relevant content may also be published by groups with no explicit scientific or academic mission. Data published by diverse commercial, academic, government, and cultural organizations are aimed at different consumers (e.g. Kansa & Wilde, 2008). Each group may publish data relevant to multidisciplinary science, but each may primarily serve constituencies favoring different data sharing standards.

Data diversity also exists due to significant variability in the form of raw data. In animal behavior, for example, "data" for a single study can consist of morphological measurements, genetic sequences, video captures, chemical swabs, audio files, etc. These diverse data types resist reduction into simple spreadsheets and lose much of their value when they are only shared as processed numeric values. In these cases, we need to find ways to allow for semantic integration among datasets composed of very different types of information.

In many disciplines, the loss of valuable and frequently non-replicable data is a crisis, leading to recommendations to openly publish available data without delay (Whitlock et al., 2010). On the other hand, we are already experiencing a data deluge. For example, the US government is an increasingly important data provider relevant to many scientific disciplines. As part of the Obama Administration's "Open Government Initiative," Data.gov recently began publishing many geospatial, demographic, environmental, economic, and other datasets. Data.gov will increasingly offer such data as dynamic services (e.g. Wilde et al., 2009a, b), but standards for sharing data at Data.gov will be determined by scientific needs and also by government and commercial interests. Blurry disciplinary boundaries and the need for multidisciplinary investigations to draw on a heterogeneous mix of collections result in a need to explore strategies for supporting scientific research where there is little prospect for consensus on common semantic data standards.

Many different research communities, including well financed national centers focused on the small sciences (e.g., NCEAS, NESCent[1], etc), share the belief that lack of archiving standards (ontological, technical, and social) is a major impediment to data sharing and interoperability. As such, people have expended extraordinary effort to create detailed metadata and other publishing standards under the presumption that shared data will not be useful if they lack such information. However, such efforts are time-consuming and costly. Sometimes these efforts can be antagonistic to the diversity of research methodologies in the small sciences; indeed, there is evidence that too much emphasis on global ontologies (even within seemingly homogenous disciplinary boundaries) can impede research and incur high costs (e.g., Menzies, 1999; Hepp, 2007).

Given these realities, we propose that we need a new frame for understanding the challenges for sharing data in the small sciences. Not only is a universal metadata standard unlikely, it is also not necessarily even desirable. Instead, we recommend "plain web" approaches for data sharing, based on loose coupling of datasets through simple services. This more feasible approach will not only enable multidisciplinary scholarship and value-extraction from shared data, but also will do so in a manner that does not incur high costs and does not impede the natural diversity of research approaches that are integral to small science domains.

4 THE PLAIN WEB AND MINIMAL REQUIREMENTS FOR SMALL-SCIENCE DATA SHARING

Small science researchers often lack the technical capacity or support needed for some of the more sophisticated cyberinfrastructure and Semantic Web frameworks deployed in better-financed domains (see Lui et al. 2007). Unfortunately, many Semantic Web implementation approaches are conceptually difficult and require mastery of a whole set of technologies (e.g., RDF, SPARQL, OWL) that have limited adoption outside a few specialized systems (Wilde 2008). Because of this gap, there is a need for a data dissemination approach aligned to the public World Wide Web. The idea of the "Plain Web" (Wilde, 2008) emphasizes using the simplest and most widely known and supported technology for any given task. In keeping with this principle, we describe the potential for using feed-based dissemination of scientific data to support third-party applications. This approach builds on the most widely used technologies on the Internet today: HTTP for service access, Atom for the service interface, and XML for the data provided by the service. This choice of widely supported technologies makes service access and consumption as open and easy as possible, as is appropriate for the diffuse disciplinary boundaries and technical constraints typical of small sciences. The Plain Web provides a pragmatic and immediately feasible foundation for researchers, as well as a useful starting point for others developing sophisticated linked data and Semantic Web services (see Battle & Benson, 2008; Greaves & Mika, 2008).

Our discussion sets aside the complex and difficult ethical, privacy, and security issues surrounding the management of sensitive data, especially datasets relating to human subjects. Such sensitive datasets will no doubt require much more complex and expensive infrastructure for ethically responsible data sharing and archiving. Fortunately, many areas of scientific investigation (ecology, geography, archaeology) have less complex privacy requirements. For these datasets, accessibility is much more of an ethical and public interest imperative. Accessibility refers to four crucial attributes: visibility, discoverability, portability, and legality. Data must be viewable by the public in order to be useful, which means that they need to be published openly on the Web and available to anyone. Data must also be published in a manner that enables their discoverability, such as by major search engines or other archival, library, or other referral systems such as discipline-specific community portals. Discoverability is greatly enhanced by use of feeds, as described below, because feeds facilitate the portability of data and their use in multiple contexts. Data portability not only encourages discovery, but also reuse. Technologies such as feeds can make it easier to aggregate and analyze datasets from multiple

----

[1] It is worth noting that NESCent, in partnership with the UNC Metadata Research Center and others, has launched Dryad, a digital repository for ecological and evolutionary data underlying published works. Dryad represents a significant step forward for less burdensome data archiving in the small sciences, though it does not yet offer the feed-based data discovery and management tools that we describe here.

sources. Finally, the data must be legally accessible, meaning that the data either reside in the public domain or have been licensed for use and reuse by anyone. Absent these qualities, published data will present so many barriers to their use that it is unlikely that they will see significant engagement or added value from peers or anyone else.

Appropriate technical implementation strategies need to be followed for useful and effective Plain Web approaches to data sharing. One of the most effective ways to use Plain Web approaches is to publish feed-based representations of the results of queries. This approach goes beyond the typical way feeds are implemented. Feeds usually publish a list of resources and metadata about those resources that may be available at a collection, usually for sharing updates about news, announcements, or comments. However, as described below, small-science data publishers can also use feeds to publish results of queries. In doing so, feeds can help facilitate analyses and aggregation across multiple data sources.

In using feeds, we strongly recommend the Atom Syndication Format (Nottingham & Sayre, 2005). Atom implementations apply architectural principles of "loose-coupling" involving RESTful ("Representational State Transfer") interactions and standards. RESTful design styles are favored because they can help reduce costs and complexity (Pautasso & Wilde, 2009). Reducing such barriers to entry are key requirements for small-science domains that typically receive little funding or technical support. Atom is useful in this respect because it is easy to implement and already widely supported in many applications, software libraries, and commercial services. Atom can facilitate Web dissemination of scientific data because it provides a common standards-based approach for disseminating custom "payloads" of XML content expressed in arbitrary schemas. In that sense, Atom offers a "standard container" for transmitting resources, including those created as a result of queries. Using Atom, resources can be shared in a way that enables complex data-structures (special purpose XML or RDF representations) to be communicated. Moreover, because Atom feeds require use of some Dublin Core-based metadata (Kunze & Baker, 2007), recipients of Atom feeds will gain at least basic metadata about the content even if they cannot fully process more specialized XML or RDF payloads. At a minimum, Atom expression of query results provides a standards-based approach for sharing URIs of resources returned from queries. Dereferencing those URIs provides opportunities to further parse the linked entity, either at the time of acquisition or in the future. As discussed below, obtaining URIs to resources in this manner can be useful for linking resources across different collections of researcher datasets, even if specialized domain or project-specific XML or RDF cannot be processed.

5 MANAGING AND QUERYING DIVERSE DATA: ATOM AND INTEGRATING "SLICES"ONCLUSION

A major driver for shared standards is the need to enable researchers to relate "slices" of data from different collections. In other words, it is more likely that specific subsets of data will be of interest to downstream users, not the entirety of a given dataset. This is especially true when relating subsets of data from different sources. Thus, users will need mechanisms to precisely identify potentially related subsets of data in different collections, and then obtain those subsets in formats useful for continued analyses.

A key requirement is to identify strategies to relate hundreds or even thousands of resources that make up different slices of data. Most current techniques (such as user-generated tagging) that help users establish relationships between web resources from different sources will not work in this scenario. User-generated tagging may not be feasible for large slices of data because it typically requires users to tag each resource individually. Some automated means will be required to relate many resources that make up large slices of data.

Atom can provide a simple and effective means to help automate annotations of large slices of data. Atom extensions for "feed-paging and archiving" enable Atom to serialize query results of any size. Thus, paging extensions represent a key requirement for using Atom to share query results. An Atom feed representation of a given slice of collection data can be read to extract URIs of individual resources that are members of that slice. By offering machine-readable lists of URIs, Atom makes it possible to rapidly add useful metadata to large slices of collections data. This approach builds on the OpenSearch standard (OpenSearch.org). Metadata created in this manner can take the form of simple "folksonomy" (unstructured) tags. Alternatively, metadata can take the form of more formalized and structured variants including domain ontologies (e.g., Abel, 2008; Gruber, 2008; Ankolekar et al., 2008) that underlie linked data systems. Regardless of the desired metadata structure, Atom expression of query results can facilitate (semi)automated approaches to metadata creation.

The approach advocated here helps bridge the gap between the Plain Web and the Semantic Web. To reiterate, URIs allow users and applications to identify meaningful units of information, even without understanding the specific models and representation formats used. Simple ways to share meaningful lists of URIs, such as a list of resources sharing some properties (as identified by queries), can be very valuable. Atom represents a simple, low barrier-to-entry approach for machine-readable expression of URIs of resources within a slice of a collection. Moreover, Atom feeds help researchers better manage the dynamic nature of collections and slices of collections. Collections typically change over time, as museums, archives, field studies, and re-analyses generate new materials, collect new data, and fix errors. Unfortunately, it is often difficult to know when a collection was last changed and how it was changed. Because an Atom feed is a service required to express publication and update time information, it helps communicate changes and updates to a collection. If Atom feeds are provided for slices of a collection, a given feed can be checked periodically (and automatically via feed-readers) to see if the slice has changed. This notification feature can be important for services that add semantically enriched metadata to slices of collections (see below) and work in a variety of social networking and instructional course applications. If the original set of resources that make up a collection slice has changed, then a user may be notified to decide whether and how to deal with the change.

Finally, this approach helps keep a human researcher "in the loop" to determine exactly how different datasets and subsets of data should be related. For many scientific applications, a researcher's domain knowledge may play an important role in resolving ambiguities in understanding a given dataset or, by extension, multiple datasets (see Palmer & Craigin, 2008). Researcher judgment is needed to determine how to compose queries that slice collections into analytically meaningful subsets of resources. A key advantage of this approach is that dataset mapping does not preclude alternatives. Any researcher would be free to select any slice of collection of interest, and apply any ontology that may be useful in relating those data with other datasets. This approach helps ensure that researchers have a choice in selecting the ontologies that best meet their needs.

6 A MASHUP APPROACH TO SCIENCE DATA SHARING

The "Programmable Web", a website that tracks developments in Web services and APIs, lists some 1700 APIs and 4700 "mashups" that make use of these APIs. This tremendous level of activity can serve as a model for making scientific data available via similar, low-barrier-to-entry means. By publishing data openly, and by making queries available as feeds, we believe that this approach helps to resolve a number of long-standing challenges that have plagued efforts to publish and share data, especially those that have focused on development of more detailed and comprehensive ontologies in the name of maximizing dataset interoperability.

First, we have to recognize the open-ended nature of possible re-uses of shared data. Rather than developing a priori and complicated standards oriented to a specific discipline or area of interest, it makes more sense to publish data more comprehensively in a manner that enables anyone with a potential interest to runs queries and engage more deeply. Second, we have to be mindful of the metadata-specification load on the owners of the original data. There are currently few incentives for researchers to invest substantial time in publishing their data, let alone publishing those data under the constraints of highly detailed and potentially complicated standards. Third, researchers should not be expected to be experts in anything more than whatever was required for them to gather and share their original data. Too often, adherence to a global ontological standard requires researchers to utilize terminology different from what they would normally choose, and to make judgment calls regarding categorization of their data that they are uncomfortable making. Data contributors should be able to contribute their data in a manner that is no more burdensome than whatever level of detail they needed for their own research purposes.

On a more technical level, the use of Atom helps to alleviate some requirements for parsing and processing system or domain-specific dialects of XML, JSON, or RDF. The number of management and visualization tools that leverage Atom continues to grow, and an expanding pool of accessible and syndicated data guarantees that such tools will quickly get more numerous and powerful. Note we are not arguing against the usefulness of either Semantic Web technologies or more detailed domain specific standards and ontologies. Such technologies and standards are very important and useful, especially within the context of a given domain or sub-discipline. However, when working across different domains or in other situations where common semantic and technical standards are hard to apply, using Atom in the way we have described is a useful and low-cost route for sharing data. The loose coupling enabled by Atom feeds of data slices is the first step in identifying potentially suitable pools of data (slices), after which researchers may want to further interrogate these data through the application of different domain-specific ontologies. Thus, by making scientific data available for serendipitous aggregation in "mashups", we can provide a foundation for more formal approaches toward data integration.

It is worth remembering that the goal in providing data for re-use is to encourage research and inquiry, not to supply ready-made research outcomes. We should expect that anyone using existing datasets will need to spend some time considering whether and how different data might relate, and then perform due diligence on the different points of intersection like any other research effort. There is a danger with over-specified standards in people assuming too much about the quality and form of the data. Yes, we want the data to be sufficiently described to be useful and merit research interest, but we should not presume that data are effectively useless if they are not already mapped to an elaborate domain ontology. Instead, we argue that datasets can still be scientifically useful if they can at least be queried in such a way as to facilitate mapping to one or more ontologies. In fact, choice in ontologies should be an important requirement for scientific data sharing. A given ontology may not be the best choice for addressing a specific research question, and investigators should be free to select those ontologies that best meet their needs (see also Pike & Gahegan, 2007; Boast et al., 2007). A key advantage in expressing query-defined slices of data as Atom feeds lies in its ability to facilitate such open ended semantic mappings.

7 DYNAMISM AND DEVELOPING FUTURE CASE STUDIES

It will take time to develop case studies to evaluate the effectiveness and research impact of the Plain Web data-dissemination methods discussed here. Currently, few scientific data publishers offer Atom feeds of query results. Those scientific data publishers that do offer Web services typically implement more complex WS/SOAP or Semantic Web services. Typically, such services are more difficult to leverage in resource-poor small-science contexts.

Nonetheless, public data already exist to make exploration of scientific possibilities feasible. For example, Open Context (http://opencontext.org), a data publication system focusing on field archaeology, already offers openly licensed datasets that can be queried to return results as Atom feeds. In addition, Dryad (http://datadryad.org) publishes many datasets relating to evolutionary and ecological sciences. In examining these two data repositories, there are many datasets that can prove scientifically useful when related. For instance, a Dryad dataset about artiodactyl (even-toed ungulates) life history, created by Price and Gittleman (2007), can be related to a dataset of artiodactyl bone specimens (identified from several archaeological sites) in Open Context. Combining the life history and archaeological data may enable certain types of scientific inquiry that bridge the archaeological past and the present and help illuminate ancient subsistence patterns and environmental conditions.

In keeping with this example, the most feasible way to currently relate the animal-bone data from Open Context to the artiodactyl life-history data in Dryad is to first download the raw tables and spreadsheets from Dryad and Open Context. A user would then relate these data using desktop spreadsheet or database applications. In most such cases, some data will need to be transformed or deleted in order to usefully merge the different datasets. However, much of that intellectual work will be lost if confined to a desktop environment. Ideally, the specific queries and other steps required to relate two (or more) different datasets should be open to inspection and reuse. In that sense, scientific advances expected by open-data advocates would be better served if datasets where not only available as static objects for download, but were also accessible via dynamic services that respond to queries.

We use the term "dynamic" services because that helps capture a crucial aspect of services valuable to data sharing. Data delivered via services means more than access to content. Rather, dynamic services deliver data together with "back-end" processes that add value to those data. These may include data curation processes, as is the case with support from a data archive; for example, the California Digital Library offers Open Context a number of data curation services exposed through software interfaces over the Web. Other back-end processes may include software-mediated responses to queries that filter or summarize a given dataset or collection. For example, Open Context provides a service summarizing the overall characteristics of its content via its faceted search application (see Hearst , 2006 for general definition and discussion). By providing this summarized information, users gain a greater understanding of a collection as a whole (Kansa & Kansa in press; Jeffrey et al., 2009). This service requires Open Context to process data "on the fly." If Open Context only offered static datasets (downloadable from a fileserver), it could not offer this kind of summarization service, or enable users to select and filter for the specific data that matches their interests. Thus, there are important advantages to be gained in moving beyond Web access to data, but also making the data available via Web-based services that help with data processing and manipulation.

Publishing data in dynamic services also helps to make sure that downstream users of those services gain access to updated data. While an individual dataset may be relatively static and "frozen" at publication, the data publishing service itself will likely continue to publish more and more data. Over time, the service will offer a more comprehensive, nuanced, and up-to-date picture that may aid subsequent interpretations.

While making scientific datasets accessible in dynamic services is certainly more challenging than making them available only for download, it is not too high of a barrier even for small-science domains. Open Context already demonstrates such services for archaeology. In addition, Google and now Microsoft have developed an impressive array of standards, protocols, and services for managing and dynamically querying shared tabular data. Both Google (with GData) and Microsoft (with OData) have embraced Atom as a foundation for building dynamic services supporting Web-based interactions with tabular data. Similarly, Yahoo has promoted another Atom-based standard, DataRSS for publishing structured data on the Web. Some aspects of these commercial frameworks may prove useful for scientific applications. While scientists also need data archiving and citation features not present in current commercial services, scientific data publishers should leverage commercial Web service developments whenever possible, either using them directly or by referencing these protocols for the deployment of their own services. The interest of these commercial giants also suggests that scientific data publishers need to look beyond Semantic Web technologies (RDF, OWL, SPARQL, etc.) and also consider syndication-based approaches to exposing data. Ideally, scientific data publishers should support a variety of Plain Web and Semantic Web methods for interacting with their data.

Finally, this dynamism opens a new door for scholarship, allowing some researchers to make scholarly contributions "as a service." Traditional artifacts of scholarly production are static objects, such as papers, monographs, or more rarely, a published dataset. However, as dynamic services increase in importance, activities related to designing, selecting, manipulating, and choreographing across these services will become an important area of research. In that sense, some important scholarly contributions will not take the form of a static article or book, but rather a service that dynamically responds to different requests. In some disciplines this trend is already impacting researcher practice. For instance, bio-medical fields are extremely fast-paced and see volumes of publication far in excess of what can be read by practitioners. To better manage this information deluge, software agents are becoming increasingly important "audiences" for biomedical publications (see reviews by Seringhaus & Gerstein, 2007; Markel, 2009). These data and text-mining software agents are the outcomes of significant research and scholarly investment. The agents power dynamic services that respond to user (and machine) queries and return useful but ever changing results. There is an inherently dynamic nature of these research products, since these products actively generate responses from changing queries and expanding collections.

8 FUTURE WORK

While the exchange of query-defined "slices" or query-defined sets of URIs can make it easier to work across multiple collections, important questions remain. In small-science domains, many datasets will be collected using different methods and research designs. Comparing "apples" to "oranges" is a major challenge in using disparate datasets from different research designs. While this is an inherent issue for using pooled data in the small sciences, the different levels of granularity in presenting and sharing these data can exacerbate the problem. Research is even more difficult if one must compare individual apples with crates of oranges! In other words, different levels of granularity complicate data comparisons and reuse (see also Riccardo et al., 2009).

In many cases, datasets are published only in aggregate, where many different scientific observations coexist in the same document, often as a spreadsheet or data table. While aggregate data-tables offer convenience for presentation and retrieval of predefined sets of data, individual records in data tables are harder to reference and link to alternative assemblages and structures. In other words, without URIs for individual units of observation, it is difficult for third parties to assemble alternative "slices" (see above) of resources and link those to other resources. Thus, granularity concerns should be a key design factor in shaping the "semantic scope" of URI-identified Web resources. We believe that small-science practitioners would benefit from continued exploration of how best to align the semantic scope of Web resources into analytically meaningful units.

Finally, relying upon query-defined slices of data as Atom feeds raises a host of query design issues. A user's options in defining slices of data will be limited by the capabilities of a collection's querying service. In other words, the semantics of a slice of data will be defined by the semantics of the query that generated that slice. To give a simple example, a given collection may only support keyword queries. In such cases, there may be a great deal of semantic ambiguity to the results of keyword queries. Even if such results were expressed as Atom feeds, they would be difficult to meaningfully relate to resources in other collections. Thus, collections that support more analytically precise queries (including support for Boolean expressions) will be more useful for the data-sharing approaches explored by this paper. Querying services should follow RESTful design principles, especially using the HTTP/GET method for retrieval of resources, and PUT, POST, and DELETE methods for the creation, modification and deletion of resources. For these later methods, the Atom Publishing Protocol complements the Atom Syndication Format-based approaches to information retrieval (Gregorio & de Hora, 2007). Common standards for query parameters may also prove useful in reducing the costs and complexity of aggregating data from multiple collections. For example, certain draft extensions to the OpenSearch protocol describe standards for geospatial and chronological queries. In addition, the OData protocol offers a more comprehensive set of specifications for describing how collections may be queried.

9 CONCLUSION

We are in the midst of a rapid shift toward Web-based research. Because many "small science" research agendas span so many disciplinary boundaries and straddle the academic, commercial, and public sectors, convergence upon complex semantic standards remains only a remote possibility. Nevertheless, there are important research questions and practical needs that demand effective strategies for working across datasets published by these disparate actors. Effective and practical methods allowing researchers to work across collections and disciplinary boundaries can have a major impact. By making datasets open to the same sort of interactions that support flourishing "mashup" efforts on the Web, we believe that scientific advances promised by open data advocates can best be realized

10 ACKNOWLEDGEMENTS

Current funding for Open Context development comes jointly from the National Endowment for the Humanities and the Institute of Museum and Library Services. Past funding comes from the William and Flora Hewlett Foundation and Doris and Donald Fisher. Thanks go to Dr. Sarah Kansa, lead editor of Open Context, for her review and feedback on drafts of this paper.

11 REFERENCES

Abel, Fabian. (2008) “The benefit of additional semantics in folksonomy systems.” Pp. 49-56 in *Proceeding of the 2nd PhD workshop on Information and knowledge management*. Napa Valley, California, USA: ACM http://portal.acm.org/citation.cfm?id=1458550.1458560 (Accessed December 13, 2009).

Albertoni, Riccardo, Elena Camossi, Monica De Martino, Franca Giannini, and Marina Monti. (2009) “Context Enabled Semantic Granularity.” Pp. 682-688 in *Knowledge-Based Intelligent Information and Engineering Systems*. http://dx.doi.org/10.1007/978-3-540-85565-1_84 (Accessed December 13, 2009).

Ankolekar, Anupriya, Markus Krötzsch, Thanh Tran, and Denny Vrandecic. (2008) “The two cultures: Mashing up Web 2.0 and the Semantic Web.” *Web Semantics: Science, Services and Agents on the World Wide Web* 6:70-75.

Baru, Chaitanya. (2007) “Sharing and caring of eScience data.” *International Journal on Digital Libraries* 7:113-116.

Battle, Robert, and Edward Benson. (2008) “Bridging the semantic Web and Web 2.0 with Representational State Transfer (REST).” *Web Semantics: Science, Services and Agents on the World Wide Web* 6:61-69.

Beers, Pieter J., and Pieter W. G. Bots. (2009) “Eliciting conceptual models to support interdisciplinary research.” *Journal of Information Science* 35:259-278.

Bennett, Daniel, and Adam Harvey. (2009) *Publishing Open Government Data*. World Wide Web Consortium (W3C) http://www.w3.org/TR/2009/WD-gov-data-20090908/ (Accessed September 9, 2009).

Blackwell, Christopher, and Thomas R. Martin. (2009) “Technology, Collaboration, and Undergraduate Research.” 3. http://www.digitalhumanities.org/dhq/vol/3/1/000024.html (Accessed December 16, 2009).

Boast, Robin, Michael Bravo, and Ramesh Srinivasan. (2007) “Return to Babel: Emergent Diversity, Digital Resources, and Local Knowledge.” *The Information Society* 23:395.

Borgman, Christine, Jillian Wallis, and Noel Enyedy. (2006) “Building Digital Libraries for Scientific Data: An

Exploratory Study of Data Practices in Habitat Ecology." Pp. 170-183 in *Research and Advanced Technology for Digital Libraries*. http://dx.doi.org/10.1007/11863878_15 (Accessed December 13, 2009).

Borgman, Christine, Jillian Wallis, and Noel Enyedy. (2007) "Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries." *International Journal on Digital Libraries* 7:17-30.

Bumpus, Hermon C. (1899) "The elimination of the unfit as illustrated by the introduced sparrow, Passer domesticus." in *Biological lectures delivered at the Marine Biological Laboratory of Wood's Holl*. Ginn & Co.

Costello, Paul. (2009) "Motivating Online Publication of Data." *BioScience* 59:418–427.

Foster, MW, and RR Sharp. (2007) "Share and share alike: Deciding how to distribute the scientific and social benefits of genomic data." *Nature Review Genetics* 8:633-639.

Fry, Jenny, and Sanna Talja. (2007) "The intellectual and social organization of academic fields and the shaping of digital resources." *Journal of Information Science* 33:115-133.

Gemmis, Marco de, Pasquale Lops, Giovanni Semeraro, and Pierpaolo Basile. (2008) "Integrating tags in a semantic content-based recommender." Pp. 163-170 in *Proceedings of the 2008 ACM conference on Recommender systems*. Lausanne, Switzerland: ACM http://portal.acm.org/citation.cfm?id=1454036&dl=GUIDE&coll=GUIDE&CFID=65784330&CFTOKEN=79244012 (Accessed November 30, 2009).

Gregorio, J, and B de Hora. (2007) "RFC 5023 - The Atom Publishing Protocol." http://www.ietf.org/rfc/rfc4287.txt (Accessed December 16, 2009).

Griffiths, Aaron. (2009) "The Publication of Research Data: Researcher Attitudes and Behaviour." *International Journal of Digital Curation* 4. http://www.ijdc.net/index.php/ijdc/article/view/101 (Accessed September 16, 2009).

Gruber, Tom. (2008) "Collective knowledge systems: Where the Social Web meets the Semantic Web." *Web Semantics: Science, Services and Agents on the World Wide Web* 6:4-13.

Hajjem, C., S. Harnad, and Y. Gingras. (2006) "Ten-Year Cross-Disciplinary Comparison of the Growth of Open Access and How it Increases Research Citation Impact." *cs/0606079*. http://arxiv.org/abs/cs/0606079 (Accessed December 16, 2009).

Hearst, Marti. (2006) "Clustering versus Faceted Categories for Information Exploration." *Communications of the ACM* 49:59-61.

Heimeriks, Gaston, Peter van den Besselaar, and Koen Frenken. (2008) "Digital disciplinary differences: An analysis of computer-mediated science and `Mode 2' knowledge production." *Research Policy* 37:1602-1615.

Hepp, Martin. (2007) "Possible Ontologies: How Reality Constrains the Development of Relevant Ontologies." *IEEE Internet Computing* 11:90-96.

Huang, Andrea Wei-Ching, and Tyng-Ruey Chuang. (2009) "Social tagging, online communication, and Peircean semiotics: a conceptual framework." *Journal of Information Science* 35:340-357.

Huynh, David, Stefano Mazzocchi, and David Karger. (2007) "Piggy Bank: Experience the Semantic Web inside your web browser." *Web Semantics: Science, Services and Agents on the World Wide Web* 5:16-27.

Jaiswal, Anuj R., C. Lee Giles, Prasenjit Mitra, and James Z. Wang. (2006) "An architecture for creating collaborative semantically capable scientific data sharing infrastructures." Pp. 75-82 in *Proceedings of the 8th annual ACM international workshop on Web information and data management*. Arlington, Virginia, USA: ACM http://portal.acm.org/citation.cfm?id=1183550.1183566 (Accessed December 13, 2009).

Jeffrey, S. et al. (2009) "The Archaeotools project: faceted classification and natural language processing in an archaeological context." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367:2507-2519.

Jung, Jason. (2009) “Using evolution strategy for cooperative focused crawling on semantic web.” *Neural Computing & Applications* 18:213-221.

Kajikawa, Yuya, Koji Abe, and Suguru Noda. (2006) “Filling the gap between researchers studying different materials and different methods: a proposal for structured keywords.” *Journal of Information Science* 32:511-524.

Kansa, Eric C and Sarah Whitcher Kansa. (In press) “ ’Mashable’ Heritage: Formats, Licenses and the Allure of Openness”, Pp. 105-112 in *Heritage in the Digital Era*, Multi-Science Publishers, London.

Kansa, Eric, and Erik Wilde. (2008) “Tourism, Peer Production, and Location-Based Service Design.” in *Proceedings of the 2008 IEEE International Conference on Services Computing, Honolulu, Hawaii*. IEEE Computer Society.

Kansa, Eric C. (2005) “A community approach to data integration: Authorship and building meaningful links across diverse archaeological data sets.” *Geosphere* 1:97-109.

Kansa, Eric C., Jason Schultz, and Ahrash N. Bissel. (2005) “Protecting Traditional Knowledge and Expanding Access to Scientific Data: Juxtaposing Intellectual Property Agendas via a “Some Rights Reserved” Model.” *International Journal of Cultural Property* 12:285-314.

Kansa, Sarah Whitcher, Eric C. Kansa, and Jason M. Schultz. (2007) “An Open Context for Near Eastern Archaeology.” *Near Eastern Archaeology* 70:188-194.

Kim, Hak Lae, Simon Scerri, John G. Breslin, Stefan Decker, and Hong Gee Kim. (2008) “The state of the art in tag ontologies: a semantic model for tagging and folksonomies.” Pp. 128-137 in *Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications*. Berlin, Germany: Dublin Core Metadata Initiative http://portal.acm.org/citation.cfm?id=1503431 (Accessed November 17, 2009).

Kintigh, Keith, W. (2006) “The Promise and Challenge of Archaeological Data Integration.” *American Antiquity* 71:567-578.

Kosala, Raymond, and Hendrik Blockeel. (2000) “Web mining research: a survey.” *SIGKDD Explor. Newsl.* 2:1-15.

Kunze, J, and T Baker. (2007) *The Dublin Core Metadata Element Set*. Internet Engineering Task Force http://www.ietf.org/rfc/rfc5013.txt (Accessed September 17, 2009).

Li Ding et al. (2005) “Search on the Semantic Web.” *Computer* 38:62-69.

Liu, Yong, Jim Myers, Barbara Minsker, and Joe Futrelle. (2007) “Leveraging Web 2.0 technologies in a Cyberenvironment for Observatory-centric Environmental Research.” Chapel Hill, North Carolina: Semantic Grid Group (SEM-RG) http://www.semanticgrid.org/OGF/ogf19/Liu.pdf (Accessed July 28, 2008).

Markel S. (2009) “BioLINK Special Interest Group Session on the Future of Scientific Publishing”. *PLoS Comput Biol* 5(5): e1000398. doi:10.1371/journal.pcbi.1000398

Menzies, Tim. (1999) “Cost benefits of ontologies.” *Intelligence* 10:26-32.

Mikroyannidis, A. (2007) “Toward a Social Semantic Web.” *Computer* 40:113-115.

Motta, Enrico. (2006) “Knowledge Publishing and Access on the Semantic Web: A Sociotechnological Analysis.” *IEEE Intelligent Systems* 21:88-90.

Nelson, Bryn. (2009) “Data sharing: Empty archives.” *Nature* 461:160-163.

Nottingham, M, and R Sayre. (2005) “RFC 4287 - The Atom Syndication Format.”
http://www.ietf.org/rfc/rfc4287.txt (Accessed December 16, 2009).

Onsrud, Harlan, and James Campbell. (2007) “Big Opportunities in Access to "Small Science" Data.” *Data Science Journal* 6:OD58-OD66.

Palmer, Carole L., and Melissa H. Cragin. (2008) “Scholarship and disciplinary practices.” *Annual Review of Information Science and Technology* 42:163-212.

Paterson, Andrew. (2003) “The design and development of a social science data warehouse: A case study of the Human Resources Development Data Warehouse Project of the Human Sciences Research Council, South Africa.” *Data Science Journal* 2:12-24.

Pautasso, Cesare, and Erik Wilde. (2009) “Why is the Web Loosely Coupled? A Multi-Faceted Metric for Service Design.” Madrid, Spain http://dret.net/netdret/publications#pau09a (Accessed February 12, 2009).

Pierce, Marlin E., Geoffrey Fox, Huapeng Yuan, and Yu Deng. (2007) “Cyberinfrastructure and Web 2.0.” http://grids.ucs.indiana.edu/ptliupages/publications/Web20ChapterFinal.pdf (Accessed April 10, 2009).

Piwowar, Heather A., Roger S. Day, and Douglas B. Fridsma. (2007) “Sharing Detailed Research Data Is Associated with Increased Citation Rate.” *PLoS One* 2:e308.

Price, S.A., Gittleman, J.L. (2007). Data from: Hunting to Extinction: Biology and Regional Economy Influence Extinction Risk and the Impact of Hunting in Artiodactyls. *Dryad Digital Repository.* http://hdl.handle.net/10255/dryad.82 (Accessed May 14, 2010).

Richards, Julian. (2004) “Online Archives.” *Internet Archaeology*.
http://intarch.ac.uk/journal/issue15/richards_index.html (Accessed March 18, 2008).

Schloen, J. David. (2001) “Archaeological Data Models and Web Publication Using XML.” *Computers and the Humanities* 35:123-152.

Schreiber, Guus et al. (2008) “Semantic annotation and search of cultural-heritage collections: The MultimediaN E-Culture demonstrator.” *Web Semantics: Science, Services and Agents on the World Wide Web* 6:243-249.

Seringhaus, Michael, and Mark Gerstein. (2007) “Publishing perishing? Towards tomorrow's information architecture.” *BMC Bioinformatics* 8:17.

Snow, Dean R. et al. (2006) “Cybertools and Archaeology.” *Science* 311:958-959.

Specia, Lucia, and Enrico Motta. (2007) “Integrating Folksonomies with the Semantic Web.” Pp. 624-639 in *The Semantic Web: Research and Applications*. http://dx.doi.org/10.1007/978-3-540-72667-8_44 (Accessed November 17, 2009).

Stronza, Amanda. (2001) “Anthropology of Tourism: Forging New Ground for Ecotourism and Other Alternatives.” *Annual Review of Anthropology* 30:261-283.

Stroup, Donna F. et al. (2000) “Meta-analysis of Observational Studies in Epidemiology: A Proposal for Reporting.” *JAMA* 283:2008-2012.

Tombros, Anastasios, Ian Ruthven, and Joemon M. Jose. (2005) “How users assess web pages for information seeking.” *J. Am. Soc. Inf. Sci. Technol.* 56:327-344.

Uhlir, Paul F., and Peter Schröder. (2007) “Open Data for Global Science.” *Data Science Journal* 6:OD36-OD53.

Uren, V. et al. (2006) “Semantic annotation for knowledge management: Requirements and a survey of the state of the art.” *Web Semantics* 4:14-28.

Whitlock, Michael C., Mark A. McPeek, Mark D. Rausher, Loren Rieseberg, and Allen J. Moore. (2010) “Data Archiving.” *The American Naturalist* 175:29-38.

Wilde, Erik. (2008) “The Plain Web.” Beijing, China: WWW2008 http://dret.net/netdret/docs/wilde-wsw2008-plain-web.pdf (Accessed May 28, 2008).

Wilde, Erik, Eric C. Kansa, and Raymond Yee. (2009a) “Proposed Guideline Clarifications for American Recovery and Reinvestment Act of 2009.” *School of Information*. http://repositories.cdlib.org/ischool/2009-029/ (Accessed March 20, 2009).

Wilde, Erik, Eric C. Kansa, and Raymond Yee. (2009b) “Web Services for Recovery.gov.” *UC Berkeley, School of Information Technical Reports* http://escholarship.org/uc/item/0fv601z8.

Wu, Xian, Lei Zhang, and Yong Yu. (2006) “Exploring social annotations for the semantic web.” Pp. 417-426 in *Proceedings of the 15th international conference on World Wide Web*. Edinburgh, Scotland: ACM http://portal.acm.org/citation.cfm?id=1135777.1135839 (Accessed November 17, 2009).

Xiao, Ruliang, and Shengqun Tang. (2008) “Towards a smallest scale of context in ontology integration.” *Wuhan University Journal of Natural Sciences* 13:407-411.

(Article history: Received 26 April 2010, Accepted 08 May 2010, Available online 29 June 2010)