

# Atom Feeds and Incremental Semantic Annotation of Archaeological Collections

Eric C. Kansa<sup>1</sup>, Tom Elliott<sup>2</sup>, Sebastian Heath<sup>3</sup>, Sean Gillies<sup>2</sup>

<sup>1</sup> UC Berkeley, School of Information

<sup>2</sup> Institute for the Study of the Ancient World, NYU

<sup>3</sup> American Numismatic Society

*ekansa@ischool.berkeley.edu*

## 1. Introduction

Archaeological and museum collections see growing online publication. However, most dissemination still takes the form of “destination websites” with little support for interoperability or aggregation. Most discussion about collections interoperability has focused on ontology considerations (Doerr and Iorizzo 2008). While these discussions are important, the conceptual challenges of sophisticated semantic integration, together with implementation complexity, create difficult hurdles. These hurdles may inhibit many from opening their collections to wider aggregation and integration.

To address this issue, this paper will outline conceptually and technically simple Atom-syndication based approaches that can open collections to enable wider aggregation. We will highlight examples from currently deployed and developing systems, including: Open Context (<http://opencontext.org>), Nomisma (<http://nomisma.org/>), and Concordia (<http://atlantides.org/trac/concordia/>). The paper will present a method enabling users to establish meaningfully semantic linkages across these systems without requiring the developers of these systems to implement a common ontology or deploy still challenging RDF/SPARQL technologies.

## 2. The Problem Area

Sharing open data can promote broader perspectives and collaborative analysis while offering new research opportunities in “small science” domains like archaeology where researchers often work individually or in small teams. As is typical of the small sciences, archaeologists often use their own models, recording systems, and taxonomies .

The diversity of archaeologically-relevant data collections is more than a function of indifference toward standards, availability of funding, and technical capacity. The wide scope of archaeological interests and disciplinary inputs also drives data diversity. Projects like Nomisma.org and Pleiades have expertly collected and maintained numismatic and historical geography collections. But these are outside the traditional scope of “field archaeology.” Not only are there different disciplinary interests and world-views to consider, but 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. Each group may publish data relevant to multidisciplinary science, but each may primarily serve constituencies favoring different needs and data sharing requirements (Onsrud and Campbell 2007).

To better work within these constraints, data sharing approaches must have very low costs and complexity. Ideally, working across different collections should not depend upon the development and implementation of a universal semantic standard. In other words, we should find ways to pursue research goals that draw upon different collections without having to solve all the semantic issues in their integration all at once. Instead, we argue that opening data in ways that facilitate incremental application of common ontologies, targeted to address specific research needs, offers a more feasible near term approach.

## 3. The Approach

By adopting a few simple design patterns described in this paper, collections systems can more easily support incremental semantic integration. The approaches described here avoid difficult semantic for-

malisms and avoid still largely unfamiliar Semantic Web technologies (RDF/SPARQL). Instead, this paper focuses on data dissemination closely aligned to the mainstream or “plain” design patterns of the public World Wide Web. “Plain Web” design emphasizes use of the simplest and most widely known and supported technology for any given task (Wilde 2008). In keeping with this principle, this paper explores feed-based (see below) dissemination of archaeological 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 appropriate for the diffuse disciplinary boundaries and technical constraints typical of archaeology. The Plain Web provides a pragmatic and immediately feasible foundation for future researchers to develop sophisticated linked data and Semantic Web services.

This pragmatic architecture can enable incremental data integration in archaeology. By incremental, we emphasize integration of disparate collections not just for the sake of integration, but also to address specific research questions relevant to the discipline. This paper demonstrates how simple, Plain Web approaches can enable researchers to integrate diverse datasets related to trade and exchange in the Ancient Near East and Mediterranean. Other than sharing some common architecture features, these collections share no common semantic standard. In fact, as discussed by this paper, Web architecture issues assume equal significance to ontologies in addressing trade and exchange topics.

#### 4. Methods

The essence of our approach is to build special purpose indexes of resources scattered in disparate collections. These indexes will enable retrieval and analysis of collections data to address disciplinary questions about trade. For this approach to succeed, two simple requirements must be met:

1. Query results as Atom Feeds: The myriad collections relevant to many archaeological research agendas can be queried or “sliced” into smaller subsets of resources depending on the organizational particulars of these collections. Researchers working with these collections sometimes need to retrieve individual resources (reports, descriptions of specific artifacts, etc.). In other cases, researchers have more interest in defining and working with entire subsets or “slices” of a collection selected through queries. In such

instances, researchers may seek to understand a slice as a whole, through statistical analysis or some form of visualization. A machine readable format needs to communicate the content of these slices in order for slices to be related across multiple collections. For this purpose, we propose using paged Atom feeds. If collections provide paged Atom feeds of their query results, third parties can easily reference subsets of these collections that meet some analytically important criteria defined by users.

2. Special purpose indexes: The second component is a third-party created service that stores an index of resources according to a common vocabulary. Once would populate data in this third party data store by obtaining URIs of resources from Atom feed expressed “slices” of collections. In our case, Concordia’s vocabulary (see below) will be used to describe these resources in ways useful for studying ancient trade and exchange.

Paged Atom feeds make it easier to relate large slices of hundreds or even thousands of resources making up a researcher-defined slice of a collection. Most techniques of user-generated tagging are not feasible for large slices of data, because user generated tagging typically requires users to tag each individual resource one at a time. Instead, some research applications require addition of metadata to many items making up large slices of collections. Currently, most Web-based collections make only human-readable representations of query result slices. This limitation makes it difficult to add metadata about each item in a slice. Fortunately, an increasing number of relevant collections offer slices of data in the syndication formats. Examples include the New York Public Library’s Digital Gallery, the UK-based Portable Antiquities Scheme, and the Flickr Commons.

Atom is useful because it is easy to implement and widely supported with many applications, software libraries, and commercial services. An Atom feed representation of a given slice of collections 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 easy 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 standard “folksonomy” (unstructured) tags, or more structured variants (see also Gruber 2008) needed for linked data systems.

To ground this discussion with an archaeological research example, we will focus on trade and exchange questions. Many online datasets, such as

Catalhoyuk and Open Context contain descriptions of objects likely to be imported from other locations. The Nomisma.org collection describes many coins found at particular locations, even though they were minted elsewhere. Because these systems offer query features, it is possible for users to define slices of these collections that may meaningfully relate to evidence for trade and exchange. For example, a user of Open Context may select a slice of data describing artifacts made with certain raw materials suggesting their import. Similarly Nomisma.org users may select slices of data about coins minted at a certain location. If these collections share query defined slices in Atom, then the Atom feed can be parsed and every item in the slice can be “tagged” using a standard vocabulary developed by the Concordia Project (sponsored by the Institute for the Study of the Ancient World at New York University). As part of this annotation, other “linked data” providers such as Geonames, Freebase, and Pleiades can be referenced to disambiguate geographic locales relevant to trade.

In our example, Condordia will store these special-purpose trade and exchange annotations of resources from Open Context, Nomisma.org and other collections. Application of this common vocabulary will allow users of these archaeological data to assert, discover and reuse the concepts of “findspot” and “origin” in existing datasets, user defined queries of those datasets, and individual objects published on the Web. Researchers can thus explore archaeologically important research questions about ancient trade and exchange using multiple collections, even though these collections themselves lack common semantic standards.

## 5. Granularity and Semantic Scope

While the exchange of query-defined “slices” or query-defined sets of URIs can make it easier to work across multiple collections, important questions about granularity remain. Comparing “apples” with “oranges” is a major challenge in using disparate datasets. 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.

In many cases, datasets are published only in aggregate, where many different archaeological 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 these to other resources. Thus, granularity concerns should be a key design factor in shaping the “semantic scope” of URI identified Web resources. We suggest continued exploration of how best to align the semantic scope of Web resources to analytically meaningful units (see also Isaksen et al. 2009b).

## 6. Conclusions

Because archaeology spans so many disciplinary boundaries and straddles 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.

Nothing about the approach advocated by this paper locks in place or “hard-codes” how one works across different collections. For many archaeological applications, a researcher's domain knowledge may play an important role in how to resolve ambiguities in understanding a given dataset, or by extension, multiple datasets (see Palmer and Craigin 2008). Researcher judgments play an important role in determining how to slice collections into analytically meaningful subsets of resources. Furthermore, in our example, the Concordia vocabulary is very simple and straightforward. Thus, archaeologists need not depend upon data modeling specialists to create relations across collections on their behalf. We hope that this simple and readily achievable demonstration of working across different collections will help motivate greater openness among collections publishers.

## References

- DOERR, MARTIN, and DOLORES IORIZZO. 2008. “The dream of a global knowledge network- A new approach.” *Journal on Computing and Cultural Heritage* 1(1): 1-23.
- 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(1): 4-13.
- ISAKSEN, LEIF, KIRK MARTINEZ, AND GRAEME EARL. 2009. “Archaeology, formality & the CIDOC CRM.” Available at: <http://eprints.soton.ac.uk/69707/> [Accessed January 11, 2010].

ISAKSEN, LEIF et al. 2009b. "Linking Archaeological Data." Available at: <http://eprints.ecs.soton.ac.uk/18240/> [Accessed December 12, 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(1): 163-212.

WILDE, ERIK. 2008. "The Plain Web." Understanding Web Evolution: A Prerequisite for Web Science. Proceedings of the Web Science Workshop of the 2008 World Wide Web Conference, Beijing, China. Available at: <http://dret.net/netdret/docs/wilde-wsw2008-plain-web.pdf> [Accessed May 28, 2008].