Smithsonian Open Data Pilot

Overview

The Smithsonian Open Data Pilot intends to demonstrate the benefits of openly sharing Smithsonian metadata with open knowledge platforms to make it more accessible and usable to people around the world. The pilot project partners representing art, science, and history divisions, as well as galleries, museums, archives, and libraries at the Smithsonian, propose an Open Data Pilot with Wikidata as the target platform as a proof-of-concept for a larger Smithsonian Open Data initiative to more widely diffuse Smithsonian collections intellectual capital around the world.

Wikidata, a five-year old project of the Wikimedia Foundation, is a free and open knowledge base that can be read and edited by both humans and machines. It is increasingly incorporated into major search engines and home assistance devices such as Alexa and Google Home. The Wikimedia Foundation is making a concerted effort to partner with news, education, and cultural heritage organizations worldwide to create a trusted and robust open knowledge backbone to enhance the Semantic Web. The Smithsonian invests heavily in its collections metadata and can have a major impact on this endeavor, as well as gain new connections and enhancements to its own data.

Project Goals & Components

Principles

The Smithsonian Open Data Pilot will adhere to the following principles:

- **Open** – In order to participate in the largest number of projects possible, collections metadata will be released as CC0 with no restriction on reuse.
- **Linked** – Metadata will be served as Linked Data which allows for richer connections to other repositories and data commons.
- **Updatable** – As the Smithsonian’s metadata is always changing, consumers should have the most up-to-date version through an “always-on” API.
- **Machine-readable** – Publishing the metadata in a machine-readable format allows for it to be used for computational analysis, and enables more facile reproduction and distribution.
- **Visitor-focused** – The deliverable should be created in consultation with end-users, and ultimately serve their needs.

Deliverables

Phase I

The project seeks to develop rich links across the participating units with two target data sets:

- **Art** – Starting with the work already accomplished by the American Art Collaborative LOD project in which SI successfully participated, we propose to develop a data set related to artists and artworks to contribute to Wikipedia’s Sum of All Paintings worldwide initiative. Participating Smithsonian units would include Archives of American Art, Freer Sackler, National Portrait Gallery, Smithsonian American Art Museum, Smithsonian Libraries.
- **Science** – The Smithsonian has already done tremendous work with data publication in the sciences; for example the Biodiversity Heritage Library Consortium (BHL) aggregates taxonomic literature across the world, and the National Museum of Natural History (NMNH) publishes discipline-specific data sets to multiple aggregators. However, this information is not incorporated into the larger World Wide Web. The target data set for this sub-project will integrate scientific specimens, their collectors, and their locations. We propose developing a data set that links:
  - the approximately 292,000 NMNH type specimens (the original specimen on which the description and name of a new species is based), with their type citations (the bibliographic reference for the original publication),
  - the persistent identifier(s) of the citation author(s) (such as ORCID, if available, which uniquely identifies academic and scientific authors),
  - the taxonomic literature in BHL (which includes the full literature referring to those specimens in many instances), and
  - field books from Smithsonian Libraries and Smithsonian Institution Archives (that often record the collecting of those very specimens).

To enable richer collaboration with Wikidata, the project will include the following deliverables for Phase I:

1. A Linked Open Dataset representing these portions of Smithsonian’s public collections metadata;
2. Effective publication of the datasets and relevant taxonomies for consumption by all interested audiences under CC0 Terms of Use;
3. Data models for integration of our target data sets with Wikidata;
4. Documented, reusable workflows using open software tools for data analysis, matching data to Wikidata identifiers, and rendering data into Linked Open Data triples.

Phase II

1. Smithsonian data publication on Wikidata;
2. Process map for new data validation and ingestion into Smithsonian collections records;
3. Test enhancements of trusted data to enhance Smithsonian collections metadata;
4. Key Performance Indicators (KPI) for performance of Smithsonian data in Wikidata.

Target Audiences

- Users of major search engines and home assistance devices;
- Cultural heritage aggregators (DPLA, Europeana);
- Open knowledge platforms (Wikipedia, Creative Commons, etc.);
- Digital humanities scholars;
- Science researchers;
Participating Units

The following Smithsonian divisions, representing the full GLAM (galleries, libraries, archives, museums) spectrum, have agreed to participate in this pilot by providing collection record data sets and making time available for review of formatted data and linkages.

- Archives of American Art
- Freer Gallery of Art and Arthur M. Sackler Gallery
- National Museum of American History
- National Museum of Natural History
- National Portrait Gallery
- Smithsonian American Art Museum
- Smithsonian Institution Archives
- Smithsonian Libraries

Additionally, the Smithsonian's Office of the Chief Information Officer (OCIO) Research Computing group has agreed to lend technical expertise and guidance during the project.

Resources Needed

Various tools, staff members, and contractors are needed to make this project a success:

- Project Sponsors (Deron Burba, OCIO, Beth Stern, Research Computing, OCIO, and 1-2 TBD Senior-Level Staff in Castle or Units)
- Project Owner (Effie Kapsalis, SIA)
- Technical Lead (Adam Soroka, Research Computing, OCIO)
- Two Wikipedians(datans)-in-Residence to coordinate units and OCIO in developing deliverables above, one heading up the art dataset and the other, science (needs funding)
- Eight Points-of-Contact from each participating unit
- Two Domain Experts (Art, Science)

Open Data and The Smithsonian

The Smithsonian, as the world’s largest museum complex with collections and scholarship across multiple disciplines, is uniquely positioned to make a sizable contribution to open knowledge and the semantic web. The Smithsonian has taken the first important steps towards unifying collections across disciplines with its Enterprise Data Access Network (EDAN) search index, e.g. Collections Search Center, which allows the public a one-stop place to search collections across the Smithsonian’s galleries, libraries, archives, museums, and the zoo online. However (appropriately for its use case) EDAN is a closed system that is disconnected from other extramural repositories. Additionally, EDAN “flattens” data, losing the important links between people, places, and events represented in the collections.

There is a grander vision the Smithsonian can pursue with its collections and associated data; by empowering scientists, researchers, scholars, and computer scientists to use the Smithsonian’s data as they desire, to create new modes of analysis, mashups, visualizations, and global networks of rapidly evolving artificial intelligence. An increasing number of people are interested in studying and using the data held by cultural heritage organizations, whether at computer science programs in universities, or open knowledge projects like Creative Commons and Wikipedia, or from specific scholarly disciplines with traditional relationships to cultural heritage organizations. There are vast numbers of machines and powerful software tools available to ingest large datasets and perform computational analysis.

Additionally, while the Smithsonian has much to give to open knowledge systems, it also stands to gain tremendously from them. A concrete example is the Smithsonian American Art Museum’s (SAAM) participation in the American Art Collaborative which has the goal of aggregating Linked Open Data about American art across American institutions. As an experiment, once SAAM published its Linked Open Data, they invited Wikidata volunteers to ingest it into the Wikidata graph database. Biographical information harvested from Wikipedia and other data sources made clear that Wikidata was a source of information that SAAM records otherwise lacked, such as an artist’s gender and race. The Smithsonian studies its collections with a specific viewpoint, and since it’s our responsibility to collect information about the collections in the Smithsonian’s context, the data can and should be enhanced by other verified sources, increasing our ability to reach new audiences.

Humanities

While the idea of using computers to aid in the study of humanities has been around since the 1940s, it is really only in the last decade that galleries, libraries, archives, and museums (GLAM) have been able to contribute more robustly to this exploding field. In addition to interest from digital humanities scholars, there are well-established efforts within the field of cultural heritage to aggregate collections to facilitate research and discovery; Digital Public Library of America for the United States, and Europeana for the European Union. Cultural heritage organizations are some of the most trusted institutions, more so than other non-profits, government agencies, and even newspapers. It is incumbent on the cultural heritage sector to participate in, and contribute to the larger knowledge ecosystem growing worldwide.

Sciences

The sciences have relied on collections-based research (biology disciplines, paleobiology, mineral sciences, etc.) powered by computers since the 1960s. Many museums including The National Museum of Natural History (NMNH) have provided online collections access since the early 1990’s. However, it was the early 2000s that saw a dramatic shift in the utility of online resources. Until that time, the large quantity of information and associated multimedia required for research had greatly outpaced the available network speeds. Since then, research has shifted to rely more and more on online available collections content, especially for large scale studies not previously possible (so called “big data”).
While in-person collections-based research remains vital for those who have or can get physical access to collections, any scientist who wants to use physical or digital collections first need to know what a museum has. In recent years, the push has been to integrate, collaborate and collectively improve the available knowledge held by scientific organizations. These globally aggregated collections could be leveraged to become even more powerful as research and analytical tools through the implementation of Linked Open Data (LOD). LOD enhances existing information, records connections of any kind between resources and provides additional valuable context, allowing for more and more informed insights.

Opportunities for the Smithsonian with the Semantic Web

Today, the Semantic Web has exponentially increased the ease of finding and navigating information online. For example, when searching for information about movies, we find that previews, showtimes, and directions are readily available without leaving the search page. When searching for tickets for air or train passage, information about flights, prices, and more is integrated right into results displays. When ‘googling’ a person or concept, semantic data from Wikipedia and other sources aggregated in Google’s massive (but closed) knowledge graph is displayed prominently on the right-side of the search results. The Smithsonian must provide semantic access to its information to be a readily-available digital-first resource for the world.

[Google Search Displaying Semantic Film Data]
More collaborative data linking and sharing will enable the Smithsonian to reach more audiences and allow it to more fully connect its content with a global user base to solve critical issues. Additionally, the enhancements made to our own data using insights gleaned from links to collegiate organizations will give us a deeper understanding of our own holdings, upon which we can base administrative decisions and priorities.

About Wikidata

Wikidata is a collaboratively edited knowledge base operated by the Wikimedia Foundation. Wikidata is intended to provide structured, semantically-enabled data to Wikimedia projects such as Wikipedia and Wikimedia Commons, and can be used by anyone else as it is under a public domain licence. Data from Wikidata is displayed in the Knowledge Panel that is featured prominently in Google searches and is increasingly becoming a foundation for both Wikimedia content on Wikipedias and in other websites.
Wikidata links to cultural heritage taxonomies like VIAF, WorldCat, as well as news, university, and schemas from other knowledge sectors. The Wikimedia Foundation is adding staff in their organization to focus on Wikidata, and the Wikimedia Foundation considers cultural heritage to be one of their key partners. As part of investing in the application of Wikidata to heritage collections, the Sloan Foundation funded a $3 million grant to support the application of Wikidata to describing materials from collections stored on Wikimedia Commons, alongside other multimedia of many types.

References

"Stewarding the Invisible: Setting the stage for Institution-wide digital preservation at the Smithsonian", pg 64