Reconciliation of Linked Data in the Cultural Heritage Sector
Notes on a Meeting Convened by the Andrew W. Mellon Foundation

Dr. Robert Sanderson, Stanford University
Date: 2014-12-17

Executive Summary

Cultural heritage institutions would benefit from the joint development of a platform for reconciliation, the process by which multiple URIs for the same entity can be related. As it is typically cheaper and easier to create a new URI for an entity being described than to discover and re-use existing ones, this is what most institutions currently do. Entities that would benefit from entity reconciliation include people, organizations, places, time periods, events, concepts and cultural heritage intellectual works and physical objects. It could be attempted from an entity-centric approach, where for example all people are reconciled, or a subdomain-centric approach, where all types of entity within a smaller set of linked resources, for example music, are reconciled. Benefits would accrue to the users in terms of access to and discovery of more relevant information, to the institutions in terms of efficiency and prestige, and to the community for education and engagement. The technology to be used should be distributed rather than centralized, and flexible enough to cope with different use cases and requirements for source data, precision and workflows. Next steps for the community include stand-alone demonstrations of the added value to their data once reconciled, better understanding of the mechanics of integrating reconciliation into our workflows, and a discussion on the domain’s unique selling points and requirements that only we can solve. Proposed actions included taking steps to address both the prospective and retrospective needs for reconciling data, and considering models, workflows and tools for current, future, and past needs.

Introduction

The Andrew W. Mellon Foundation convened a meeting on the topic of reconciliation of Linked Data on December 1st and 2nd, 2014. The meeting was attended by many of the leading institutions in the use of Linked Data across the cultural heritage sector, including representatives from several national libraries, museums, a national archives, major research libraries, large scale aggregators and invited technical experts. For two days, the group discussed the topic from a variety of perspectives, commencing with scene-setting presentations and concluding with a round-table discussion about priorities and appropriate next steps that the community can take.
**Reconciliation?**

The Mellon Foundation has supported a substantial number of internally successful projects that have adopted, implemented and promoted Linked Data. These projects have created billions of URIs identifying people, places, objects, concepts and much more. However, the value of Linked Data is in the overall network of connected resources, not in the greatness of any individual institution or system. While there are the best intentions to re-use resources, time pressures and other factors make the expedient route to simply create new identifiers for entities that may already have been identified and described elsewhere in the network, and thus not capitalizing on Linked Data’s inherent potential.

The notion of reconciliation is, in essence, the process by which the multiple identities assigned to the same entity are discovered and related together. By reconciling the URIs, the descriptive information associated with them can be merged to provide richer and deeper knowledge than any single institution can hope to provide alone. This process provides a linking mechanism after the fact; retroactively converting the stand-alone identifiers into providing useful anchoring points into the larger network. The more URIs are reconciled, the larger the network, giving more information and context to make further reconciliation processes easier and more accurate.

**Causes for URI Proliferation**

In order to understand reconciliation, it is important to understand the causes of the situation that it aims to correct: the proliferation of multiple URIs that identify the same entity. The basis of any network is that the nodes are individually valuable, and that linking to the node is cheaper than recreating it. It is practically free to link to any major news site on the internet, and recreating the content in that site would be enormously expensive. That many other sites link to that site makes it easier to find, either by manually following links or using any modern web search engine. This is typically not the case for many of the entities described by cultural heritage institutions using Linked Data.

Libraries, for example, have a great deal of overlap in their holdings. It was reported that Stanford and Cornell University Libraries have approximately 50% of their holdings in common; both universities have millions of the same books. This pattern was expected to continue for all other major research institutions. This means that the libraries have a great deal of overlap in the people, such as authors, referenced in their bibliographic records, in the corporate bodies or organizations such as publishers, in the concepts, events, times and places that are the subjects of those books. Further, the bibliographic descriptions of the books are very short, not natively linked data and the scale of the data is relatively large. There are few, if any, central aggregation sites that can be programmatically used to discover the
entities, and no individual library has any great claim on being the single institution that should be responsible for minting and maintaining the URI for the works that it holds.

This results in a situation where it is very costly to discover the appropriate node in the network to link to, the existing URI for the entity, and very cheap to recreate the information that would be available if the URI was known as all institutions have very similar bibliographic descriptions. As the URIs themselves are cheap to create, if in the long term expensive to maintain, there is very little incentive to do anything other than to continue to create data silos.

This is compounded by the multiplicity of bibliographic linked data models and disagreement regarding their quality and value. Even if a URI exists and is known for the entity, but the description does not follow local practice, there may be reluctance to re-use the content. Every institution has a different definition of what constitutes the high quality content they wish their site to present to users.

The way to drive these costs down is by increasing the value of the network, at the same time making it easier and more efficient to discover the identifiers that act as anchors or hubs within that network.

Reconciliation is not a fix that will prevent this situation from continuing. It’s a method for reducing the damage to the network when institutions create new and unnecessary identities, and a way to retro-fit the network on top of those existing identifiers to attempt to increase the value of the network. If the reconciliation effort is ultimately successful, it will be a (very long) one off process. The success will be when the value of the network has made it more costly to recreate rather than to link.

**Entities to Reconcile**

There was significant agreement on the classes of entity that could profit from being reconciled. In particular, of general interest were:

- Agents
  - Persons
  - Organizations
- Places
- Periods of Time
- Events, as a combination of activity at a particular time and place
- Cultural Heritage Objects
  - Individual objects, such as museum items
  - The groupings that those individuals are intellectually part of
  - The abstract concept of the Work, as opposed to any realization of it
- Concepts or Subjects
It was noted that all of the above have challenges in terms of their definition or identity; the use of pen-names when publishing, such as Mark Twain versus Samuel L. Clemens, and the division and recombination in different forms of political boundaries and companies over time provide challenges for any notion of identity, let alone a shared understanding driven by linked data generated *en masse* from records never intended to fulfill such a purpose.

A well-received suggestion was that one institution had experienced success by instead of focusing on the class of entity, focusing on a domain of interest and reconciling the entities within it. This had two immediate and significant impacts: that experience was gained with all of the types of entity on a manageable scale; and that a connected network was more likely to form, as there are more intra-domain references than inter-, and the presence of all the types of entity gives more opportunities for those relationships to be discovered rather than only person to person relationships, for example.

Another suggestion was focusing on entities that are sufficiently well defined, such as a single physical object, rather than the eternally arguable definitions of purely intellectual constructions such as a work or concept.

**Benefits of Reconciliation**

Especially in the domain of cultural heritage, no single institution could ever hope or expect to have all of the information relevant to a particular information need, yet users, both researcher or interested lay person alike, want to have a single point of access to as much information as possible. Reconciliation of URIs allows the network to provide the additional information, rather than every institution needing to create it all. It enables the bulk enrichment of cultural heritage descriptions across all sources of information that use the reconciled URIs or contribute data to the reconciliation service.

The benefits at a more granular level fall into three classes, based on their position in the information usage lifecycle.

The end user of a service with reconciled URIs has more direct access to information about the entities she is interested in via a single service. This results in better efficiency for her research, as rather than having to manually search multiple datasets, the data is already brought together for her via linked data. Furthermore, the discovery and navigation steps to get to the data are likely to be faster, easier and more reliable, because the service has access to more information about the entities. As research becomes increasingly inter-disciplinary, the graph of information that spans multiple institutions and datasets is more likely to be of use than single silos. There are also possibilities for new end user interfaces based on the reconciled graph of entities, rather than single records.
The institution gains benefits in the efficiency and quality of cataloging or description of the entities of interest. By reconciling existing identifiers, and using other institutions’ content, the amount of time spent describing shared entities can be reduced sometimes to zero. This saves time and thus reduces the cost, or allows the time to be spent on the institution’s unique entities to provide more value back into the network. The ability to provide richer descriptions of local entities brings prestige to the institution when others reuse the data, as well as providing a better estimate of value for the collection: holdings that are shared broadly are perhaps less valuable than unique items. This has additional implications for prioritization of digitization, the sorts of items that should be highlighted in exhibits, and so forth. It was noted that the efficiency argument alone does not encourage the publication of data back to the network, but instead to cut costs and be solely a consumer. The value of the ability to measure the impact of publishing data towards prestige factors, improved scholarship, teaching and learning should not be underestimated.

Thirdly, the community as a whole benefits from reconciliation. By providing richer descriptions of our unique holdings, and otherwise working together to provide accurate and detailed information about joint holdings, the community will see better engagement from all sectors. As the public and purse-holders are better educated about cultural heritage content, they will value it all the more. By generating the content jointly and not duplicating effort, we help ourselves at the same time as helping others. Reconciliation with entities in the broader web of data lets the cultural heritage sector enrich, and be enriched by, the data asserted by companies and organizations that may otherwise have no contact with the community.

The sentiment expressed was that if reconciliation is done correctly and fully integrated into our professional and institutional workflows, it will be transformative to the sector, not just another data format for description. It was noted that deeply integrating linked data and reconciliation into description represents a fundamental shift in cataloging practice.

**Features of a Reconciliation Engine**

Humans are comfortable with ambiguity, but machines are not. Humans use words and language, whereas machines rely on symbols, in this case URIs. It is especially important not to be ambiguous with information on top of which more layers are intended to be built. In the same way as building a house on sand is a bad idea, building rich services on top of ambiguous data will inevitably lead to disaster. In order to reconcile the ambiguity of the identifiers of our shared entities, a platform needs to be developed. Various features of such a platform and its development path were discussed to inform a route forwards.

The platform should be dynamic and flexible. Different types of entities will require different algorithms or network weightings to insure the accuracy of the process.
This configuration might first be done by hand, but with a sufficiently large gold standard data set available to evaluate the quality of the configuration, machine learning processes could be used to automatically discover the best approach.

The platform should also be distributed and part of the web, rather than a centralized and tightly controlled service. It should align with the broader set of standards and processes being developed, providing a framework for technical participation and innovation. The ability to contribute services and data should allow many institutions to participate, with proper recognition of their input, and even to the extent of taking crowd-sourced assertions and usage patterns into account.

As with any information system, there is always a trade-off between precision (the accuracy with which entities are linked) and recall (the number of appropriate entities that are linked). Some use cases will require high accuracy, whereas some will derive greater benefit from linking more entities together, even if some are not correct. The error rate at which the data becomes worthless will depend on the context of use, and thus the flexibility of the platform is important to ensure it fulfills the community’s varied needs. Building high precision, high recall services will require data that is fit for purpose, which in turn requires a better understanding of the purpose, context and proposed benefits of linked data in this context.

It will be hard to maintain confidence in the results, given the challenges to be overcome. The trustworthiness of the results can be demonstrated by providing the provenance of the information used and the steps taken in the reconciliation process, rather than being a black box service that gives answers of unknown quality from unknown evidence. The ability to request this information is important, but should not adversely affect production use of the service once trust has been established.

The context that surrounds an entity is important information to have access to and improves the reconciliation process. The more connected the network is around the entity, the more information is available to determine if two entities are in fact the same. At the same time, people and institutions will disagree about “the truth” and there will be conflicting assertions both about the entities and even more so when the context around them is taken into account. The platform needs to be designed to allow for disagreement, and capture the consensus plus the scope of possibilities rather than asserting a single world-view as absolute truth.

One participant, a long time discussant in the semantic web and standards, noted that while linking data and reconciliation is rife with complexity and potential difficulties around edge cases, there is also tremendous overlap and agreement around simple use cases. Productive reconciliation and linking of data can begin on the simpler end of the spectrum, and grow from there; this approach can generate manifold benefits and further momentum without being snarled in theoretical discussions.
**Future Steps**

While reconciliation might by itself promote an environment where it is no longer needed, a lot of discussion also focused on steps that could be proactively taken to slow the pace of creation of new URIs and thus push back the doomsday clock hands for the time at which there is simply too many URIs to be able to ever process.

One suggestion was that we should start with creating new data that did not need reconciliation, get workflows and systems working with that data to demonstrate the value, and only then worry about the legacy. This could be jumpstarted by assigning and publishing identifiers as early in the information lifecycle as possible, either with the publisher of the content or when the first record describing it is made. Those identifiers would need to be registered publicly, and respected internationally. Existing infrastructure should be used, such as promoting the identification of people with ORCIDs.

While this will clearly help in the future, the immense backlog of decades of data to bring into the network is a challenge that should be tackled sooner rather than later to allow for iterative development, experimentation and innovation, and to gather together as a community around a platform and best practices. As a community we must identify requirements that are both crucial to support and unique to our domain. There are many organizations working on descriptions of people or books, so where do we carve out a niche that lets us shine on the World Wide Web’s global stage? While there were suggestions, there was little consensus and this question bears further discussion in the short term.

A second question raised but not well answered was what can be done quickly that will be useful to the largest set of institutions in the sector? What changes to current workflows can we make, that will promote the visibility of the needs and justify development and adoption of the solutions? We have identifiers, too many of them even, but what is needed is tooling to get them into the business processes. A simple example given was a widget with dropdown menus supporting type-ahead or auto-complete for names of entities, that could be jointly developed and broadly used.

**Conclusion**

The meeting ended with a short and crystal clear statement, harking back to the first presentation: As a community, we need more avalanche, fewer special snowflakes.

The next steps for the community include stand-alone demonstrations of the added value of reconciliation, better understanding of the mechanics of integrating reconciliation into our workflows, and the discussion on the domain’s unique selling points and requirements that only we can solve. Both the prospective and retrospective needs for reconciling data, in terms of models, technologies and integration with current workflows, must be considered for the long term success of linked data in cultural heritage.