Thanks for having us. We’re here today to speak about our project, Drawings of the Florentine Painters, a linked catalogue for the semantic web.
The project is funded by the Samuel H. Kress Foundation, and aims to develop the first online database of Florentine Renaissance drawings. It does this by creating Linked Open Data from the three editions of Bernard Berenson’s foundational text, The Drawings of the Florentine Painters. We presented on the first phases at last year’s ARLIS/NA + VRA Conference.
And we mostly focused on the Creation and Modeling phases.
Although really, those two things happened at the same time.
And actually, a more accurate representation of how things went looks more like this. Today, we’ll give you a brief overview of the Creation and Modeling phases, a taste of the transformation phase, and the nitty gritty details about that last slice, publishing.
But first, a bit about Berenson. Bernard Berenson was an American art critic and art historian, famous for his connoisseurship of Italian Renaissance art and his role as an advisor to collectors and dealers.
The data source for this project is not one but three different editions of his famous catalogue raisonné of Florentine Renaissance drawings.
The entries describing drawings were transformed from this
...to this, by collaborators in Florence working on data entry. Later, I and other collaborators used tools like CLICK Open Refine, Python scripts, and Power Tools for Google Sheets to further transform the data and enrich it with URIs from CLICK the Getty Vocabularies, VIAF, and Geonames.
Through all of these interventions, we realized that we had actually transformed some of the data so much that it could really no longer be called data from one of the three editions; so we added another source, and called it “project data”.
The next phase was modeling
I won’t go into too much depth about modeling, since we did that in our last presentation, but I will mention that we used the CIDOC-CRM as our primary ontology, CLICK and created a few properties to extend the CIDOC-CRM to capture some of the more granular attributions Berenson used like “school of”
Here you can see a diagram of the model we created for this data. Here is a later model made using Gliffy. As I mentioned, many of the phases of this project overlapped; Modeling took quite a long time, and it was essential to do this work while cleaning and formatting the data itself. Doing so enabled patterns to emerge and inform the modeling decisions. For example, we created a more explicit recto/verso division. In the Berenson’s entries, recto and verso were indicated only if a sheet had a drawing on the front and the back; but in our dataset, we designated recto and verso as components of the larger drawing entity, thus making explicit and machine-readable a feature that was previously inferred and understood implicitly in the mind of the reader. Now all drawings have at least one described side, including elements like title and commentary associated with that side.
The diagrams were essential reference for the transformation phase of the project.
By transformation, I mean here the transformation of the data from XML format to RDF triples. We used 3M, a web-based mapping software used by the British Museum and developed by FORTH-ICS (a computer science research group based in Greece). 3M allowed us to take an XML version of our spreadsheet and produce RDF triple statements. You can see a selection of these on the left side of the screen.
3M uses XPath to create a mapping from the XML file to the entities and properties of RDF. Although our dataset is small, we didn’t do the transformation in this web-based version of 3M; we downloaded the transformation software and ran it locally using the command line. These mappings can be used for new data; for example, we recently added more images and were able to relatively quickly run them through a 3M mapping we’d made earlier.
So once we finished the transformation phase, we have our data in “triples” or RDF. In order to publish this data, we had to deploy and develop on top of a series of open-source pieces of software.
To begin with, we needed a triplestore that could serve our data to the public, and we went with a product called Blazegraph, which is robust, open-source, feature rich, and relatively easy to deploy. More importantly, it’s bundled with another product we are using called Mataphacts, which is a middleware that is being used by the ResearchSpace project at the British Museum.
Metaphacts is a platform that provides a suite of tools that enable cultural heritage institutions to host, publish, and even catalog their collections in Linked Open Data. Although the platform is still in its early stages of development, in collaboration with ResearchSpace, it’s rapidly evolving and will soon offer a more or less all-in-one solution for all of your Linked Open Data needs. Because it is still in development, we were not able to use it as a search interface, but we did use it to serve our data, and we did use some of visualization tools to be able to browse the semantic structure of our records.
Here you can see a sample of one of our records in the metaphacts platform. It uses templates to construct customized views of your records, even based on the type of record or collection. Although there is a bit of a learning curve to be able to customize the interface, it mostly just requires knowledge of HTML5, SPARQL.

With some additional development, we will also be able to fully harness the power linked open data to be able to pull in information dynamically from other collections, such as the British museum which have already publish all of their records in Linked Open Data. We have already taken advantage of this in one case by reconciling our records with those of British Museum, simply by matching up the drawing inventory numbers, and we have programmatically pulled in most of their images and made them available in our catalog.
And here we can see how the platform allows you to browse through the data, and also see it’s semantic structure.

This is particularly useful in our case because of the way that we utilize named graphs as containers for each edition and our project data. This structure provides provenance for our triples, which you can see on the bottom right of the screen next to each statement.

So for the time being it’s great for users like data scientists or digital humanist that want to explore and download our data, but it isn’t really ready for non technical users until the ResearchSpace search interface templates are complete.
With very recent developments to the metaphacts platform, we were able to begin to develop a type of cataloging interface, to be able to update the components our data that may need to be changed in the future. Here we created forms to be able to add images from external collections, as well as links to museum records. This will allow us to easily make small-scale updates to the dataset in a graphical user interface. This is a key component for us to be able to maintain our records in a single database, and will allow us to move towards a single platform for managing our collections.
Once the ResearchSpace templates are released this summer, we will have a robust contextual search system, that will allow scholars to perform searches with very specific criteria, even across datasets.
And here is a sample of what a record will look like in the metaphacts platform. However, since this ResearchSpace template is still unreleased, we felt that the next best product out there that was able to meet our needs was Project Blacklight, which is a popular open-source ruby on rails platform in use by many institutions, both as a library catalogs, but also for digital projects.
Here you can see what our catalog looks like—scholars have the ability to search and browse according to the various facets of our data, in a relatively easy to use and straightforward interface.
And here you can see what one of the results pages look like
We began the process with a mockup of how we would like one of our records to look. We wanted the graphical user interface to resemble the structure of our data, so we placed general information about each catalog record in the top right, and data that varies from one edition to another below. This also allows scholars to track and quickly visualize differences in titles and attribution through the various editions.
And here is what it looks like in the public interface.

In addition to all of the original data that was cataloged, transformed, and semantically enriched from the three editions, we also have data from a number of external vocabularies, along with links to the museum records if we were able to find them.
Since Blacklight uses a solr index as its data source, we essentially had to flatten out our richly encoded Linked Open Data so that it could be mapped to a solr index. You can find more information on this process and view the code on our GitHub repository, which we will share on our last slide.
We also wanted to give users the ability to explore the printed text, as well as view the images that were out of copyright, like the plates from the 1903 edition, in a IIIF viewer. For the printed text we were able to use our institutional repository, which uses Mirador as the viewer, but for the plates we needed to have more control over the implementation of the IIIF server, so we ended up deploying our own, and we used DigiLib for that.
The final piece of publishing is to provide documentation for the end-user so that they can understand how to use and interpret the dataset. In our project, this takes the form of the mapping diagrams we showed before, an explanation of the named graph structure, and some sample SPARQL queries. We’ll also be providing an extensive technical narrative describing the project’s processes.

Documentation

- Gifty diagrams
- Named Graph structure
- Sample SPARQL queries
- Technical narrative explaining choices made in data creation, modeling, transformation, and publishing
As for future impacts, having completed this project, it will allow us to re-use the existing infrastructure (including the blazegraph database, the metaphacts platform, and templates, the SPARQL endpoints, and the IIIF server). As our digital collections grow, using this shared infrastructure for these projects will allow us to search across collections. Our experience with this project has therefore lowered the barriers for future projects.
I’ll pass it on back to Alex to talk about some of the lessons that we’ve learned working on this project.
I think one of the main lessons learned for us was about time management and the how time- and labor-intensive data creation is. We have been creating and tweaking the data the entire time; while this process got faster since we had certain scripts already written, it did mean that we had less energy to devote to other aspects of the project.
A challenge that came up during the publishing phase was managing changing data. Because our data originated in a spreadsheet and not a database, and then the data was transformed through a variety of separate scripts, we didn’t really have a linked, unified workflow. Since things were done in pieces, it was sometimes difficult to go back a step, especially once we had our data in the triplestore. Unlike a relational database, where one can simply update a record, triplestores make it very challenging to change existing data. You have to use SPARQL to delete the existing triples and add the updated ones as new data. The new cataloging templates Lukas mentioned can help with this on a small-scale, but batch updates for us require serious SPARQLing.
Relatedly, it took us time to learn how to use 3M for mappings. Managing the data in the triplestore required skills with SPARQL, which Lukas luckily has. But budgeting time to learn these new languages and programs is important. Both Lukas and Alex attended trainings with ResearchSpace to learn about the CIDOC-CRM and how to use 3M.

In our case, it was (and is) also a challenge to ensure the skills and lessons learned could be translated into a format that would be useful to the institution after the project’s end. The major challenge here is that many of the team members are only involved on a part-time, contract basis with this grant only. Frequent phone calls and project management tools like Confluence helped with this, but with software with a high learning curve like 3M often only one person had the practice and expertise to carry out those processes.

Another issue was knowing the requirements for what we wanted our end-user application to do. We spent a lot of time wrestling with the data; for me, it felt like without having the form and structure of the data clarified, we weren’t able to articulate what we wanted for end users early enough in our project timeline. For this...
reason, we drew on talent internal to our team to build the Blacklight site; ideally, we would have been able to articulate the requirements earlier in the process and hire an outside developer for that portion of the project.
Thank you! Questions?

Project team
Jonathan Nelson, Co-Primary Investigator
Lukas Klic, Co-Primary Investigator
Alexandra Provo, Project Manager
Cristina Pattuelli, Design & Methodology
Matt Miller, Design & Methodology
Carlos Acevedo, Collaborator, Metadata Cleanup & Interlinking

Cataloging
Sara Della Bianchina
Dominic Ferrante
Kelley Garrison
Alice Parri
Diana Ingerman
Sylvia Rubin-Budick

Support
Samuel H. Kress Foundation

Contact
Alexandra Provo
aprovo@gmail.com | @AlexandraAlisa
Lukas Klic
klic@itatti.harvard.edu

Catalog: http://florentinedrawings.itatti.harvard.edu
Data: http://data.itatti.harvard.edu
Github: https://github.com/villaitatti