



Università
Ca' Foscari
Venezia

Corso di Dottorato di ricerca
in Informatica
ciclo XXXI

Tesi di Ricerca

**Digital Publishing and
Research Infrastructure for
Cultural Heritage:**
an Institutional Roadmap

SSD: INF/01

Coordinatore del Dottorato

prof. Riccardo Focardi

Supervisore

prof. Salvatore Orlando

prof. Andrea Torsello

prof. Dorit Raines

Dottorando

Lukas Klic

Matricola 956256

Table of Contents

I	Introduction	4
II	Digital Transitions	10
III	Capturing & Cleaning Collection Data	36
IV	Synergizing and Publishing Data	89
V	Visual Search for the Semantic Web	121
VI	Towards an Open and Collaborative Digital Art History	148

Index of acronyms & initialisms

AAC	American Art Collaborative
AAT	Art and Architecture Thesaurus (Getty Vocabulary)
API	Application Programming Interface
CIDOC-CRM	CIDOC object-oriented Conceptual Reference Model
CV	Computer Vision
EDM	Europeana Data Model
ETL	Extraction Transformation and Load
FADGI	Federal Agencies Digital Guidelines Initiative
FC's & FR's	Fundamental Categories and Fundamental Relationships
ICOM	International Council of Museums
IIIF	International Image Interoperability Framework
JSON-LD	JavaScript Object Notation for Linked Data
LOC	Library of Congress
ML	Machine Learning
OWL	Web Ontology Language
RDF	Resource Description Framework
SAAM	Smithsonian American Art Museum
SKOS	Simple Knowledge Organization System
SPARQL	SPARQL Protocol and RDF Query Language
TGN	Thesaurus of Geographic Names (Getty Vocabulary)
ULAN	Union List of Artist Names (Getty Vocabulary)
URI	Uniform Resource Identifier
VIAF	Virtual Information Authority File
W3C	World Wide Web consortium

Abstract

The burgeoning field of Digital Humanities has seen a great deal of interest in methodologies that support the exploration, cross-pollination, and programmatic analysis of heritage collections across the web of data. Although the heritage community has generally agreed that these data should be semantically enriched using the CIDOC Conceptual Reference Model and published as Linked Open Data, a lack of agreement at both data and infrastructural levels has hindered advancements that would allow for greater data integration and computational exploration. This project provides an institutional roadmap for publishing such data in a Semantic Web research environment, proposing a set of best practices for the community. Using a collection of 230,000 images and index metadata, this project presents methodologies and tools for data cleaning, reconciliation, enrichment, and transformation for publishing in a native Resource Description Framework system. A semantic framework for integrating computer vision services enables subsequent enrichment and visual analysis, enabling the mass-digitization of heritage collections with minimal burden on institutions, all while ensuring the long-term preservation and interoperability of these data at a global scale.

I - Introduction

The burgeoning field of Digital Humanities (DH) opens scholarship to new and exciting possibilities. Cutting across traditional disciplinary boundaries, it enables new research questions to be posed, and offers unprecedented opportunities for scholarly collaboration. The field has greatly enriched and enlivened scholarship in Art History. Major digital projects such as the Medici Archive Project¹ and the Venice Time Machine,² as well as numerous panels at major conferences are fueling a greater interest in the intersection of DH and Art Historical studies. While an initial transformation for Art History to the digital world in the early 2000's was prompted by the obsolescence of film³, and grew out of a need rather than opportunity, the recent shift is a result of technological developments that enable deeper insight into objects and historical documents at an unprecedented scale. New and exciting paths of scholarly inquiry are being tested and defined, enabling the analysis of large data sets, visualizations, geospatial mapping, network analysis, and the application of machine learning and computer vision to humanities data. While various digital initiatives in this sphere make substantial contributions to our understanding of the history of culture, the often siloed nature of Digital Humanities projects make sifting through data repositories on the open web a cumbersome process. This dissertation presents an argument for the adoption of Semantic Web and Linked Open Data (LOD)

1. *Medici Archive Project Mission | The Medici Archive Project*. <http://www.medicini.org/mission/>. Accessed 5 Feb. 2019.

2. Abbott, Alison. "The 'Time Machine' Reconstructing Ancient Venice's Social Networks." *Nature News*, vol. 546, no. 7658, June 2017, p. 341. www.nature.com, doi:[10.1038/546341a](https://doi.org/10.1038/546341a).

3. Zorich, Diane M. "Digital Art History: A Community Assessment." *Visual Resources*, vol. 29, no. 1–2, June 2013, pp. 14–21. *Taylor and Francis+NEJM*, doi:[10.1080/01973762.2013.761108](https://doi.org/10.1080/01973762.2013.761108).

technologies as two foundational components to form the backbone of DH research, enabling the storage and computational analysis of highly expressive machine-readable research data.

LOD and allows for data produced from other tools to be woven together in ways that allow them to integrate, becoming more useful at a global scale. Natural Language Processing (NLP), enables the programmatic analysis of large corpuses of texts, allowing language to be interpreted by machines and facilitate analysis. Named-Entity Recognition, enables the extraction of entities (persons, places, events, things, etc.) from free-form texts, and provides structure to otherwise unstructured data. Computer Vision opens doors to new insights and interpretation of images, allowing for functionality such as visual search, visual cataloging and image classification. Machine Learning allows for the processing, parsing, and classification of research data en-masse. Network Analysis and visualizations of matrices of structured knowledge allow for deeper insights into the history of cultural phenomena through distant reading⁴. Coupling these methodologies with the Semantic Web, published as Linked Open Data, contribute to a vibrant culture of open scholarship and collaboration among researchers, disrupting barriers posed by proprietary databases where information is kept in silos. The nature of this machine-readable data lends itself well to a more playful and serendipitous discovery, making it attractive and engaging to both undergraduates and seasoned scholars alike. The overarching aim is to advance the global paradigm shift in publishing models, away from an inward looking, closed and costly strategy, towards an open and inclusive model that encourages collaboration and open-access.

4. Moretti, Franco. *Distant Reading*. Verso, 2013.

This project presents a use case that weaves together these technologies in fruitful ways resulting in a research platform that aims to facilitate access and interpretation for a set of collections data. Although the broader project is ongoing, seeking to integrate an array of functionality including digital scholarly publishing, geospatial mapping, and Computer Vision services for the Semantic Web, the scope here will be limited to the enrichment, reconciliation, and publishing of a single collection. The resulting platform will integrate and unify various collections of digital assets and metadata, making them available to scholars in a machine-readable format, offering novel discovery tools and facilitating serendipitous discovery. Although the movement towards Linked Open Data has proliferated in recent years, with the LOD cloud doubling between 2014-2018,⁵ it is still in an early stage of adoption and the scholarly community lacks the necessary toolsets to render and facilitate the research process fruitful to scholars in the humanities seeking to leverage these advancements, in particular at the application layer. Only in 2015 did the World Wide Web consortium (W3C) publish their first set of architectural recommendations for Linked Data at the application layer, forming a proposal for how these services should interact with one another on the web⁶. By demonstrating numerous use-cases, tools for the manipulation and generation of collections data, coupled with an open-source research and discovery platform, this project demonstrates how cultural heritage data can be effectively published as LOD with a minimal burden on institutions and scholars, while maximizing the benefit to the scholarly community.

Making data available in a machine-readable format, allows researchers to interpret the source content in ways that are not possible in the print form. Unlike a traditional database where

5. *The Linked Open Data Cloud*. <https://lod-cloud.net/>. Accessed 5 Feb. 2019.

6. *Linked Data Platform 1.0*. <https://www.w3.org/TR/ldp/>. Accessed 5 Feb. 2019.

all data remains on the host's website, LOD allows researchers to have direct access to the data through open data services. Though the Web has made digital content more accessible, most researchers must still gather data by digging through collections in different repositories and sifting through an array of different databases, finding aids, and documents. The LOD initiative has emerged in recent years as a powerful set of techniques for publishing and interlinking structured data, in such a way that it can be processed by machines and openly shared on the Web. LOD technology is based on a small set of well-established and widely accepted open web standards, allowing data to connect from heterogeneous sources and make it publicly available and reusable in different contexts. The goal is to overcome the barriers posed by institutional repositories and databases, where information is kept in separate silos. LOD enables new and integrated views of cultural heritage data facilitating discovery and analysis, which in turn, leads to unanticipated research paths and the creation of new scholarship.

The very nature of research on artworks makes it especially well-suited to the LOD environment. Scholarship often focuses on identifying and articulating the complex network of relationships that surrounds a work of art. On the one hand, there are numerous types of objects: preparatory sketches, figure and composition studies, drawings, paintings, and copies of works. On the other hand, there are a range of individuals: artists, patrons, observers, collectors, scholars, and conservators. Transforming these data into LOD allows for these networks of actors and objects to be expressed in a machine-readable form that is programmatically queryable. Additionally, scholars and researchers from around the world have free and easy access to this data, which will in turn support an array of research projects in outside

applications. Finally, integration methods can enhance an existing dataset with external contextual data (including spatial and temporal data), and vice versa.

The evolution of this project is described in the forthcoming pages, detailing the methodology and choices surrounding the digitization and publishing of a collection of 230,000 images and associated metadata index that document 115,000 photographs (one image for the recto and one for the verso). The preparation of the data, reconciliation, enrichment, and transformation to LOD, together with the images serve as a testing ground for other mass-digitization efforts. Options are evaluated for publishing LOD in a native Resource Description Framework (RDF) environment that does not require continuous ETL (Extraction Transformation and Load) processes from relational databases. Finally, a framework for integrating computer vision services into the research platform is explored, enabling advanced tools for data enrichment and visual analysis. Overall, this projects aims to make a series of concrete contributions to the field:

- Provide a state of the art for the current landscape of publishing collections data as Linked Open Data
- Share methodologies and best practices for the transformation of collections data to RDF
- Provide a high-level analysis of the transformation process that describe the effort required for similar projects
- Build and share a suite of tools for cleaning and reconciling data
- Publish data and images (115 photographic records) in RDF
- Publish a dataset of provenance data that provides a snapshot of the distribution of Italian Renaissance Art in the 20th century
- Propose a semantic framework and architecture for integrating Computer Vision services into Linked Data environments

A user interface that has direct access to the source data will be provided, as well as access to data dumps and a SPARQL endpoint for advanced users. These have arguably become the gold standard for publishing data, especially for cultural heritage. In addition to making the data fully searchable, the significance of having the catalog data available as LOD lies in the possibility to connect the data to internal resources and external datasets through URIs (Uniform Resource Identifiers). For this reason, entities have been reconciled to corresponding records from external authorities, including the Getty's Union List of Artist Names (ULAN) and Art and Architecture Thesaurus (AAT), GeoNames, the Virtual Information Authority File (VIAF), and WikiData. This project will serve as a representative case study, guiding institutions and individuals seeking to publish collection or research data. DH researchers can reuse its methodology, which in turn can be applied to similar art history initiatives, especially those which focus on scholarly inventories and photographic documentation. There is a large gap in the field of publishing cultural heritage data in semantic web environments. To date, best practices for publishing and transforming such data have yet to emerge, with strong disagreements between North American and European constituents. For this reason, this dissertation addresses many of these disagreements and proposes a path forward, aiming to serve as a set of guiding principles for institutions looking to mass-digitize and publish similar collections.

II - Digital Transitions

The transition to a digital world for historical photo archives has a long and complex history spanning multiple decades. This chapter will provide a state of the art for the field, outlining current shortcomings and challenges, advocating specific solutions at the data layer as well as the infrastructural level for further elaboration in subsequent chapters. In order to address these issues it is first important to survey the landscape of Art History and their constituents, art historians.

In addition to substantial methodological shifts in theory and practice, the discipline of Art History has undergone a series of practical transformations in the past two decades. First with the transition of images away from film, and subsequently with texts, both primary and secondary sources being increasingly available in the digital form. Although the field has seen a slower transition in comparison to other humanistic disciplines,⁷ its focus on images can benefit greatly from toolsets that allow for additional insight into images. Early efforts in this digital transition have focused on providing access to large amounts of visual and textual documentation,⁸ while recent efforts have moved away from supporting digitization efforts to supporting digital tools that facilitate interpretation and the research lifecycle. This is particularly evident when observing the funding programs of large foundations, both public and private that provide financial support to large projects. Nearly all of the major foundations supporting Art

7. Kohle, Hubertus. "Kunstgeschichte und Digital Humanities. Einladung zu einer Debatte." *Zeitschrift für Kunstgeschichte*, July 2016, pp. 151–54.

8. Gahtgens, Thomas W. "Thoughts on the Digital Future of the Humanities and Art History." *Visual Resources*, vol. 29, no. 1–2, June 2013, pp. 22–25. *Crossref*, doi:[10.1080/01973762.2013.761110](https://doi.org/10.1080/01973762.2013.761110).

History and other humanistic disciplines have shifted funding strategies away from digitization efforts to supporting digital research efforts⁹, with many funders making explicit statements regarding the cessation of funding for digitization. Funders that continue to support digitization, often support collections that are either at risk or of high intellectual value, such as the Council on Library and Information Resources (CLIR).¹⁰

As a result of the shift to digital images, slide libraries with millions of slides that at one time served to provide reproductions of works of art for teaching and research purposes in research centers and universities across the world, have become obsolete in the span of just a few years. Large image databases, including ArtStor, Prometheus, Europeana, and especially Google, have rendered many of these collections and departments superfluous. The vast majority of these slides are reproductions from printed books, serving faculty in lectures and teaching environments to display works of art, and serve mostly as reference material. Today, most of these slides have little intrinsic value on their own as they are merely reproductions that can be found elsewhere in books or on the internet. Occasionally, alongside many of these slide libraries we find collections of historical photographs, the primary instrument used to study works of art by scholars during the first half of the twentieth century. These collections, which often include works of art that have been lost to wars or private collections, are rich with historical documentation, containing over a century of annotations by trusted scholars. These collections have fallen prey to the same fate as many of the slide collections that are no longer in use, as the overhead required to access and maintain this material is too high and the vast majority of

9. Zorich, Diane. *A Survey of Digital Cultural Heritage Initiatives and Their Sustainability Concerns*. Council on Library and Information Resources, 2003.

10. "Hidden Collections • CLIR." *CLIR*, <https://www.clir.org/hiddencollections/>. Accessed 6 Feb. 2019.

scholars are not willing to make the effort to explore them. There is however, a large portion of undiscovered material waiting to be found in these archives, as they go back much farther than slide collections. As eloquently stated by the former director of the Getty Research Institute “photo archives are sleeping beauties ... they are waiting to be discovered and kissed.”¹¹

Due to the trajectory of research practices in the humanities shifting increasingly towards digital resources, younger scholars are generally expecting material to be available online and are far less willing to enter a library or archive to request printed material. The lack of financial and institutional support to digitize and publish photo archive material, places these collections at risk of being left behind and forgotten. The Getty Research Institute, whose underlying mission is “dedicated to furthering knowledge and advancing understanding of the visual arts”,¹² is a prime example of such case. The Institute’s photographic archive, which houses a collection of over two million¹³ historical photographs, in addition to copies from collections around the world, has seen almost no use by scholars in recent years, with the archive staff being reduced to a single member over the past two decades. Interest in these printed collections within archives and special collections of libraries has been dwindling for quite some time. On the flip side, many collections see an increase in requests of physical archival material once collections are digitized. This kind of increase in activity is a testament the fact that these materials are of interest to scholars, but they are generally not willing to take the initiative to peruse through an archival collection that has no point of access in a digital form. Not digitizing and making the

11. ‘Photo Archives Are Sleeping Beauties.’ *Pharos Is Their Prince*. - *The New York Times*. <https://www.nytimes.com/2017/03/14/arts/design/art-history-digital-archive-museums-pharos.html>. Accessed 23 Feb. 2019.

12. *About the Research Institute* (Getty Research Institute). <http://www.getty.edu/research/institute/>. Accessed 23 Feb. 2019.

13. *Photo Archive* (Getty Research Institute). <http://www.getty.edu/research/tools/photo/>. Accessed 23 Feb. 2019.

material accessible places this material at a risk of being lost and forgotten, especially as institutional memory of these collections fades and their contents are no longer known by individuals.

In order to overcome the technical, logistical, and financial burdens of publishing material in a digital form, it is of vital importance that more institutions and individuals develop open and reproducible solutions that can facilitate and streamline these publishing efforts for cultural heritage. Sharing and reusing tools and methodologies that can enhance access and interpretation of these materials is a crucial first step in facilitating the process that will enable institutions to open and publish collections. When engaging in large-scale digitization and publishing projects, it is also important to find a balance between tasks that are manageable and provide a sufficient scholarly value. Every task that requires a substantial amount of effort should be carefully weighed to evaluate the usefulness and utility of their outcome. This project seeks to reflect on this web of decisions, making a series of proposals based on the experience publishing the historical photo archive of Villa I Tatti, the Harvard University Center for Italian Renaissance Studies.

Linked Open Data for cultural heritage

The Semantic Web, a principle first introduced by Tim-Burners Lee in his 2001 *Scientific American* article,¹⁴ acted as a prelude to the Linked Open Data (LOD) movement. While LOD is a method for publishing and enabling connections across the web of data, the Semantic Web

14. Berners-Lee, Tim, et al. "The Semantic Web: A New Form of Web Content That Is Meaningful to Computers Will Unleash a Revolution of New Possibilities." *ScientificAmerican.Com*, May 2001.

provides a layer of meaning for this information that provides context to the connections.¹⁵

Coupled, these two technologies can offer complex and intricate views of cultural heritage data. They greatly facilitate the analysis of primary and secondary source material through highly dynamic and extensible query parameters, which allow for computationally actionable reasoning on data that can be visualized in ways that can offer new perspectives on results.

Although the web has made digital content more accessible, scholarship in the digital realm is still rather two-dimensional, where researchers must sift through siloed repositories in order access data. The LOD initiative has gained much traction in recent years as a powerful tool for representing units of information that can be processed by machines. This technology is based on a set of well-established open web standards that allow data to connect from heterogeneous sources. Data flows freely across the Semantic Web environment, defined by the [World Wide Web](#) consortium as a “common framework that allows data to be shared and reused across application, enterprise, and community boundaries.” In this way, Linked Open Data contributes to a vibrant culture of open scholarship and collaboration between researchers and repositories.

There is an increasing interest in publishing cultural heritage data as Linked Open Data and in developing systems that enable the aggregation of resources found in disparate collections, creating a more seamless user access. A growing number of cultural institutions, including many art museums, are converting their metadata and vocabularies into LOD and leveraging these technologies to support new modes of discovery for artworks in a digital

15. Oldman, Dominic, et al. “Zen and the Art of Linked Data: New Strategies for a Semantic Web of Humanist Knowledge.” *A New Companion to Digital Humanities*, edited by Susan Schreibman et al., John Wiley & Sons, Ltd, 2015, pp. 251–73. *Crossref*, doi:[10.1002/9781118680605.ch18](https://doi.org/10.1002/9781118680605.ch18).

context. At a global level, Europeana collects metadata from more than 3500 cultural institutions across Europe using an upper-level ontology, the Europeana Data Model (EDM), to aggregate the metadata and enable discovery from a central access point.¹⁶ The Europeana Linked Open Data pilot dataset was first released in 2012 from a subset of more than 200 institutions. It is available in the Resource Description Framework (RDF) format, structured by EDM and accessible through dereferencing URIs, download, and a SPARQL endpoint.¹⁷ Museums like the Amsterdam Museum,¹⁸ an early adopter of the technology, and the National Museum of Finland¹⁹ have also transformed their metadata into RDF, similarly providing what have become standard access modes: dereferencing URIs, data download, and SPARQL endpoints.

The British Museum is perhaps in the vanguard of this front, as it has published the complete set of records of its Collection Online --about two million items-- as linked data using the CIDOC Conceptual Reference Model (CIDOC-CRM) ontology, and provides multiple access methods.²⁰ The richness of their collections data is unparalleled by almost any other museum collection, and the data published as LOD had been recognized as one of the most comprehensive collections of data published using the CIDOC model. The Museum's collection data is also used in the ResearchSpace project,²¹ which seeks to create a set of research tools by

16. "Reasons to Share Your Data on Europeana Collections." *Europeana Pro*, <https://pro.europeana.eu/page/reasons-to-share-your-data-on-europeana-collections>. Accessed 22 Feb. 2019.

17. Haslhofer, Bernhard, and Antoine Isaac. "Data.Europeana.Eu: The Europeana Linked Open Data Pilot." *International Conference on Dublin Core and Metadata Applications*, vol. 0, Sept. 2011, pp. 94–104.

18. de Boer, Victor, et al. "Supporting Linked Data Production for Cultural Heritage Institutes: The Amsterdam Museum Case Study." *The Semantic Web: Research and Applications*, edited by Elena Simperl et al., Springer Berlin Heidelberg, 2012, pp. 733–47.

19. Hyvönen, Eero, et al. "Linked Data Finland: A 7-Star Model and Platform for Publishing and Re-Using Linked Datasets." *The Semantic Web: ESWC 2014 Satellite Events*, edited by Valentina Presutti et al., Springer International Publishing, 2014, pp. 226–30.

20. "British Museum Publishes Its Collection Semantically." *British Museum*, <https://www.britishmuseum.org/about-us/news-and-press/press-releases/2011/semantic-web-endpoint.aspx>. Accessed 22 Feb. 2019.

21. *ResearchSpace - a Digital Wunderkammer for the Cultural Heritage Knowledge Graph*. <https://www.researchspace.org/>. Accessed 10 Sept. 2017.

leveraging their semantic relationships, using RDF datasets structured by the CIDOC-CRM. Other museum collections include the Yale Center for British Art, which publishes its collections metadata in RDF with the common goal of enriching, enhancing, and interlinking data to support repurposing and research.²² In 2014, the Smithsonian Museum of American Art (SAAM) began publishing its metadata as linked data, allowing users to explore connections and identify relationships between artists and artworks. SAAM has linked artists from their records to DBpedia as a means to connect with external linked data sources, as well as to the Getty Union List of Artist Names (ULAN) and artists in the Rijksmuseum collection.²³ This connection to the Rijksmuseum is enabled by the fact that Rijksmuseum has also made its extensive art collection data accessible via LOD technologies using the EDM ontology shared by Europeana members. As of March 2016, the Rijksmuseum has released over 22 million RDF triples describing more than 350,000 objects.²⁴ This has enabled enhanced search, discovery, and contextual information through cross-referencing, interlinking, and integration. For example, objects from the Rijksmuseum print collection are linked to related publication sources from the National Library of the Netherlands.²⁵

Publishing RDF datasets from within a single institution is an important first step, however interlinking these datasets with external sources, as exemplified by the SAAM and the Rijksmuseum, evinces the diverse potential of Linked Open Data. With this goal in mind, the

22. *Linked Open Data | Yale Center for British Art*. <https://britishart.yale.edu/collections/using-collections/technology/linked-open-data>. Accessed 10 Sep. 2017.

23. Szekeley, Pedro, et al. "Connecting the Smithsonian American Art Museum to the Linked Data Cloud." *The Semantic Web: Semantics and Big Data*, edited by Philipp Cimiano et al., Springer Berlin Heidelberg, 2013, pp. 593–607.

24. Dijkshoorn, Chris, et al. "The Rijksmuseum Collection as Linked Data." *Semantic Web*, vol. 9, no. 2, Jan. 2018, pp. 221–30. *content-iospress-com.ezp-prod1.hul.harvard.edu*, doi:[10.3233/SW-170257](https://doi.org/10.3233/SW-170257).

25. Ibid

American Art Collaborative (AAC) has formed a consortium of fourteen American art museums that seek to add the museums' collections data to the linked data cloud to “exponentially enhance the access, linking, and sharing of information about American art in a way that transcends what is currently possible with structured data”.²⁶ In addition to the SAAM and the Yale Center for British Art, the consortium includes a diverse set of institutions such as the Autry Museum of the American West, National Museum of Wildlife Art, and the Indianapolis Museum of Art. The project is now in the midst of its implementation phase to convert a “critical mass” of participating museums' metadata to LOD, but has yet to agree on many details. The Consortium for Open Research Data in the Humanities (CORDH), of which the author is a founding partner, is a similar initiative based in Europe, that seeks to harmonize data standards and infrastructures to facilitate interoperability and the cross-pollination of research data²⁷. Related to the linking and enrichment of art collections, the Getty Vocabulary Program has released their set of vocabularies as LOD including the Getty Thesaurus of Geographic Names (TGN), the Union List of Artist Names (ULAN), and the Art & Architecture Thesaurus²⁸(AAT), although they are in the process of restructuring the ontology to align more closely with recent developments in the field. In the way that these vocabularies have enabled general interoperability between institutions and collections through the standardization of metadata, they support interlinking between RDF datasets. LOD technologies continue to be adopted by an increasing number of art-

26. *About the American Art Collaborative* | American Art Collaborative. <http://americanartcollaborative.org/about/>. Accessed 10 Sep. 2017.

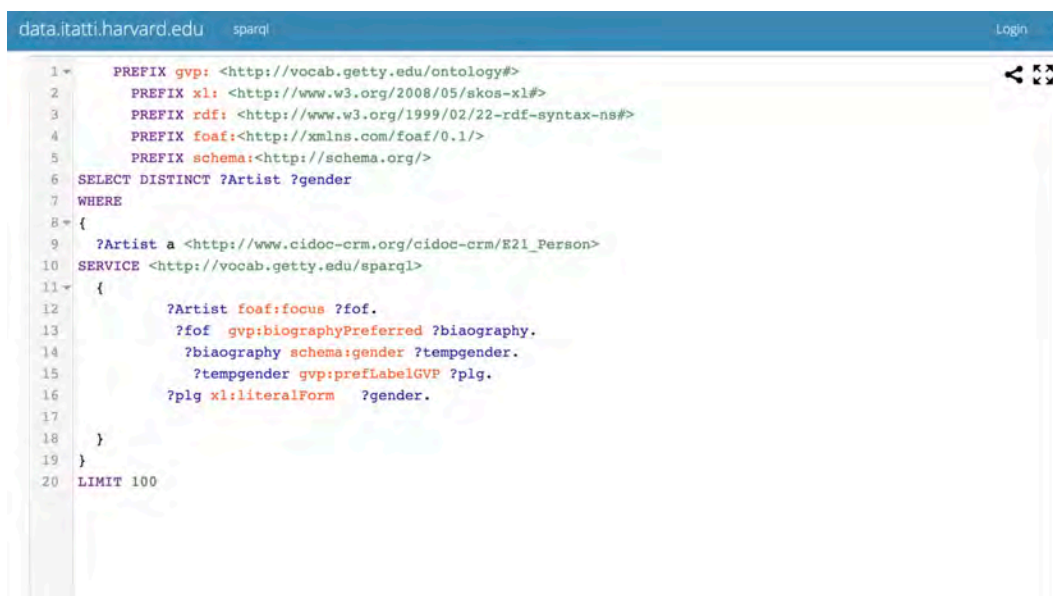
27. “Consortium for Open Research Data in the Humanities.” *Consortium for Open Research Data in the Humanities*, <https://www.cordh.net/>. Accessed 23 Feb. 2019.

28. *Getty Vocabularies*. <http://vocab.getty.edu/>. Accessed 22 Feb. 2019.

related projects and institutions, enabling new collaborative initiatives to improve methods of exploration and discovery of digital objects and information.

Photographic archives lend themselves particularly well to LOD technology due to the complexity of the objects needing to be described. Historical photographs offer a myriad of layers that require description; on one hand there is the physical object itself: its material, technique, dating, location, development process, provenance and possible annotations. On the other hand the photograph itself is depicting an object that also has a place and time of its own, a creator, medium, technique as well as an elaborate ownership history. Traditional image catalogs would attempt to capture the complexity of these objects through a traditional field-value relationship. The expressiveness of LOD allows for infinite descriptive layers, as statements are not limited to these predefined structures. Employing vocabularies and taxonomies that describe relationships and hierarchies between entities, provides additional context. The argument for moving collections to RDF-native repositories can be exemplified by numerous use-cases. For example, using a traditional catalog, place names are not contextualized, and a search for “artworks created in Tuscany” are not able to retrieve results where those search terms are not explicitly contained within the record in full text. Using a vocabulary such as Geonames, Linked Data systems can apply reason to data that is structured, understanding that Florence and Siena are part of Tuscany, and hence it should return results with artworks created in those cities. A researcher interested in the use of materials in a particular space and time, can ask these more complex questions far more efficiently with RDF datasets. Another example of the usefulness this contextual data was demonstrated in a project published by the author in 2017 for which he was principal investigator. A collection of drawings by Florentine Painters was published in a

Semantic Web environment²⁹ and a posting was shared on Facebook. One user, who was unaware of the fact that the digital publication was based on a printed book by Bernard Berenson, rightly criticized that there were no women represented in the catalog. Although gender information was not included in source data, the artist entities were linked to the Getty ULAN, so a quick SPARQL query (Figure 1) was able to take the entire set of artist records and programmatically retrieve the gender, quickly confirming an assertion that would have otherwise taken several hours to prove or disprove.



```

data.itatti.harvard.edu sparql Login
1 PREFIX gvp: <http://vocab.getty.edu/ontology#>
2 PREFIX xl: <http://www.w3.org/2008/05/skos-xl#>
3 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
4 PREFIX foaf: <http://xmlns.com/foaf/0.1/>
5 PREFIX schema: <http://schema.org/>
6 SELECT DISTINCT ?Artist ?gender
7 WHERE
8 {
9   ?Artist a <http://www.cidoc-crm.org/cidoc-crm/E21_Person>
10  SERVICE <http://vocab.getty.edu/sparql>
11  {
12    ?Artist foaf:focus ?fof.
13    ?fof gvp:biographyPreferred ?biaography.
14    ?biaography schema:gender ?tempgender.
15    ?tempgender gvp:prefLabelGVP ?plg.
16    ?plg xl:literalForm ?gender.
17  }
18 }
19 }
20 LIMIT 100

```

Figure 1: Artist gender query againsts ULAN endpoint

Beyond leveraging taxonomies and vocabularies to provide additional insight into these data, ontologies such as the CRM provide the framework of relationships between entities (people, places, things, events) in a highly expressive way that allows for additional levels of querying. With such computationally actionable data, visualizations and geospatial tools can

29. *The Drawings of the Florentine Painters*. <http://florentinedrawings.itatti.harvard.edu/>. Accessed 23 Feb. 2019.

offer a unique view of results, with the ability to discover commonalities or trends that would be otherwise very difficult or impossible to discover in traditional catalogs.

The movement towards Linked Open Data is an important one that also requires a critical mass of organizations to adopt the technology. As more institutions publish their data and make it available, it becomes easier for other institutions to interlink records and reuse data across the web. The growth of the LOD cloud can be exemplified as in figure 2 from 2014 and 3 from 2018.

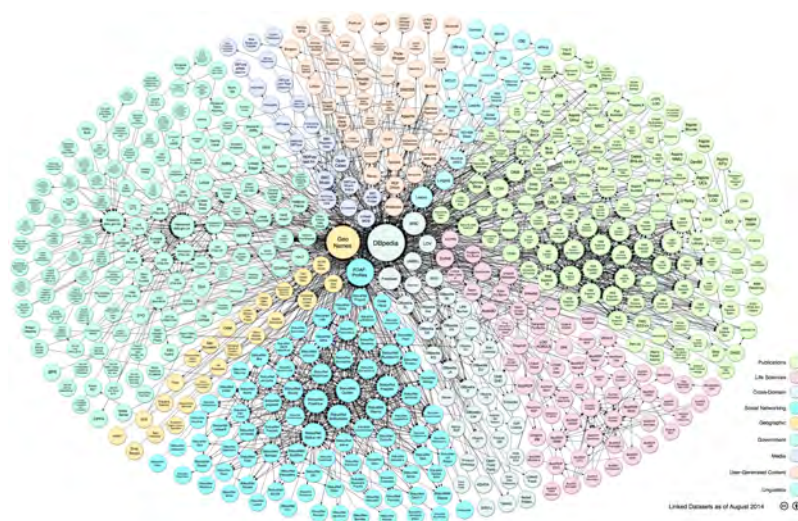


Figure 2: The LOD Cloud August 2014 (570 data sets)

Image Credit: <https://lod-cloud.net/#diagram>

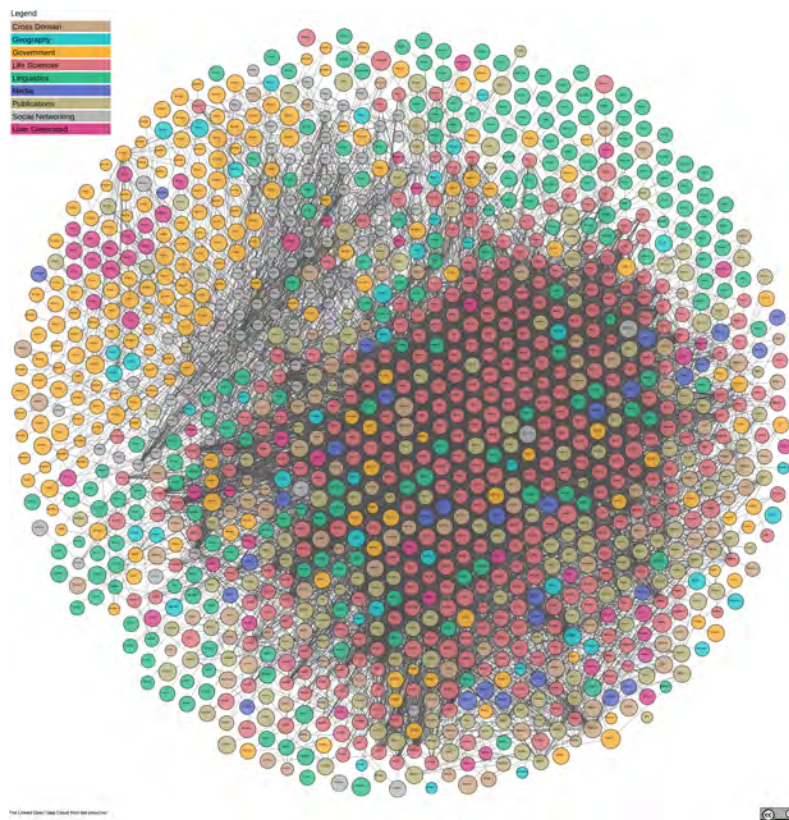


Figure 3: The LOD Cloud in July 2018 (1220 Datasets)

Image Credit: <https://lod-cloud.net/#diagram>

Although this count is not comprehensive, in these visualizations we can see that the number of data sets has more than doubled over a period of four years. This is significant for cultural heritage institutions seeking to adopt this technology, as it is evident that the community is moving steadily in this direction and is looking to adopt it as the standard. Particularly significant is the growth of datasets such as Wikidata, the Getty ULAN and AAT, Geonames, VIAF and Worldcat, that already provide identifiers and data for artists, collections, institutions, terms, places, and bibliographic entities, among others. When creating cultural heritage datasets ab ovo, Linked Data technology can be leveraged in order to avoid having to rekey or republish data that is already available in the web, by simply providing the identifier of an existing entity.

Despite the rapid growth towards the adoption of Linked Data, disagreements in the way these data are published are still present in the community and remain an obstacle to wider adoption. These disagreements range from which ontology to use, the way those ontologies are implemented in the data model, the data serialization formats, as well as at the infrastructure level.

Expressiveness vs. Interoperability

Cultural heritage publishing is currently at a crossroads in determining the best Linked Data serialization to use. The RDF³⁰ standard emerged as a W3C standard initially in 1999, and has seen a great deal of adoption. It has however recently come under much criticism for its level of complexity³¹, in particular for the existing developer community that is less familiar with graph-based data structures. As a result, the JSON-LD³² serialization emerged in 2010, with the aim of lowering the barriers to adoption and data readability. As articulated by Manu Sporny in his 2014 article, JSON-LD was born from “the desire for better Web APIs”.³³ Sporny makes several arguments in favor of JSON-LD, including the need to support lists within one’s data, a clearer and more transparent data model for developers, and data interchange that is based on APIs rather than SPARQL. The discussion generally boils down to complexity, both at the data level and the accessibility of that data. In the cultural heritage domain, the term “Linked Open

30. *RDF - Semantic Web Standards*. <https://www.w3.org/RDF/>. Accessed 23 Feb. 2019.

31. J. Rochkind. “Is the Semantic Web Still a Thing?” *Bibliographic Wilderness*, 28 Oct. 2014, <https://bibwild.wordpress.com/2014/10/28/is-the-semantic-web-still-a-thing/>.

32. *JSON-LD 1.1*. <https://www.w3.org/2018/jsonld-cg-reports/json-ld/>. Accessed 23 Feb. 2019.

33. *JSON-LD and Why I Hate the Semantic Web | The Beautiful, Tormented Machine*. <http://manu.sporny.org/2014/json-ld-origins-2/>. Accessed 23 Feb. 2019.

Usable Data” (LOUD) has been coined³⁴ by the Linked.art community, with the objective to find a balance between usability and complexity. This balance is sought by creating “baseline patterns” for entities³⁵ which are shared across the data model, employing the use of vocabularies (mostly AAT) to provide formal definitions to data, rather than leveraging the full expressiveness of the CIDOC ontology. Despite the fact that the author is on the editorial board of the Linked.art community, this project argues for the use of RDF-based systems, employing the full expressiveness of the CIDOC ontology. This position is taken based on the fact that data in the humanities is far less certain, or “fuzzy”, than that of the sciences. This fuzziness adds complexity, and that complexity is lost with data models that are less expressive. Linked.art and JSON-LD aim to simplify concepts that are inherently complex, resulting in a loss of representation as they exist in the real world, all for the sake of interoperability.

The Linked.art data model is not the only model seeking to add more structure to the cultural heritage domain. The ArtFrame ontology³⁶, a project funded by the Andrew W. Mellon foundation seeks to provide a similar structure for artworks, and is a competing ontology to the Linked.art movement. ArtFrame is a much less mature ontology that provides even less coverage than the Linked.art model, as it is largely based on Dublin Core and does not formally employ the use of specific vocabularies to add expressiveness to data. Aside from preliminary presentations given at conferences such as the Art Libraries Society of North America³⁷, and the EuropeanaTech³⁸ conference in 2018, the literature on these topics is very slim at the time of

34. *LOUD: Linked Open Usable Data*. <https://linked.art/loud/index.html>. Accessed 23 Feb. 2019.

35. *Baseline Patterns*. <https://linked.art/model/base/>. Accessed 23 Feb. 2019.

36. *ArtFrame - LD4P Public Website - DuraSpace Wiki*. <https://wiki.duraspace.org/display/LD4P/ArtFrame>. Accessed 23 Feb. 2019.

37. Billey, Amber M., et al. *The Outcome of the ArtFrame Project, a Domain-Specific BIBFRAME Exploration*. 2018. academiccommons.columbia.edu, doi:[10.7916/D8281M24](https://doi.org/10.7916/D8281M24).

38. Robert Sanderson. *EuropeanaTech Keynote: Shout It out LOUD*. <https://www.slideshare.net/azaro42/europeanatech-keynote-shout-it-out-loud>.

writing. The documentation that does exist, currently does not have a formal method for defining relationships between entities, or is limited to a single use-case.

Although Linked.art employs the use of the CIDOC ontology, it attempts to restrict its use in favor of interoperability. While the argument for interoperability is a strong one, this project advocates for a two-tiered approach where data is published first in its full expressiveness in one layer, with a second materialized data layer providing a layer for interoperability. While the Linked.art model could be used for this second layer, it is restricted to the domain of two and three-dimensional artworks. This project instead advocates for the use of Fundamental Categories and Fundamental Relationships³⁹ (FC's and FR's) as an interoperability layer between institutional repositories. This data layer addresses the recall gap when working with CIDOC-based datasets by abstracting entities into broader categories that allow for grouping and more intuitive searching. It predefines a set of categories (persons, events, places, objects, concepts) and relationships between them (thing-place, thing-actor, event-place, etc.) in order to allow for the querying of these entities at a more generalized level. Rather than traversing multiple nodes in the graph to find relationships between categories, it creates a direct link between them to allow for greater interoperability and faster searching. The advantage of using FC's and FR's over an application profile such as Linked.art, is that it is not domain-specific, allowing artwork data to be connected to that of other domains. This allows for a more generalized model that allows for example, one to make statements about the relationship between artworks and bibliographic citations, where authors are making assertions about an artwork. More details on

39. Tzompanaki, Katerina, and Martin Doerr. *Fundamental Categories and Relationships for Intuitive Querying CIDOC-CRM Based Repositories*. 2012, p. 153.

some of these modeling choices will be outlined in chapter three, where specific use-cases will be addressed.

Photo Archives Online

With the exception of the Fondazione Zeri in Bologna, we have not seen many collections from historical photo archives published as Linked Open Data.⁴⁰ We have also not seen any standards for publishing models emerge in the same way that we have for bibliographic entities. While Italy has the “Scheda F”⁴¹ that emerged from the Istituto Centrale per il Catalogo e la Documentazione, other countries have not managed to follow any kind of standard publishing model that allows for the interoperability of these data. In the United States there have been many efforts to create standards for the documentation of artworks, but these efforts have largely failed as there has not been any agreement or widespread adoption. While efforts such as Artframe have enjoyed financial support from the Mellon Foundation and institutional support from many large universities, competing efforts from the American Art Collaborative⁴² and Linked.art are overshadowing those efforts, making any kind of global movement towards standardized documentation practices all the more difficult.

The lack of agreement has created challenges for institutions seeking to publish this material, as there is no international body that can provide guidance in the community. While the

40. Gonano, Ciro Mattia, et al. “Zeri e LODE: Extracting the Zeri Photo Archive to Linked Open Data: Formalizing the Conceptual Model.” *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, IEEE Press, 2014, pp. 289–298. *ACM Digital Library*, <http://dl.acm.org/citation.cfm?id=2740769.2740820>.

41. Berardi, Elena. *NORMATIVA F - FOTOGRAFIA: STRUTTURAZIONE DEI DATI E NORME DI COMPILAZIONE*. 4.0, ISTITUTO CENTRALE PER IL CATALOGO E LA DOCUMENTAZIONE, 2015, p. 180. Zotero, <http://www.iccd.beniculturali.it/getFile.php?id=4479>.

42. *American Art Collaborative*. <http://americanartcollaborative.org/>. Accessed 23 Feb. 2019.

CIDOC community has made attempts to take on this role, until recent years they have not made much effort to disseminate results and create a more inclusive community that moves beyond Europe. This has resulted in a lack of progress at the infrastructural level, with most image catalogs running on the same or similar infrastructure as the did in the early 2000's. Although most institutions do not have the resources to digitize, let alone catalog these collections, we rarely see online catalogs that provide robust points of access, along with clean metadata and fully scanned images. Many of these databases address art documentation needs with library discovery platforms. These catalogs in most cases perform some full-text indexing, and may provide limited faceted navigation and browsing of images. The image catalog of the NYARC consortium⁴³ for example, uses the ExLibris product Primo⁴⁴. Other catalogs, such as that from the National Gallery of Art or the Fondazione Federico Zeri, are basic documentation systems with limited functionality.

43. *New York Art Resources Consortium* | <http://nyarc.org/>. Accessed 23 Feb. 2019.

44. "Primo Library Resource Discovery Solution." *Ex Libris*, <https://www.exlibrisgroup.com/products/primo-library-discovery/>. Accessed 23 Feb. 2019.

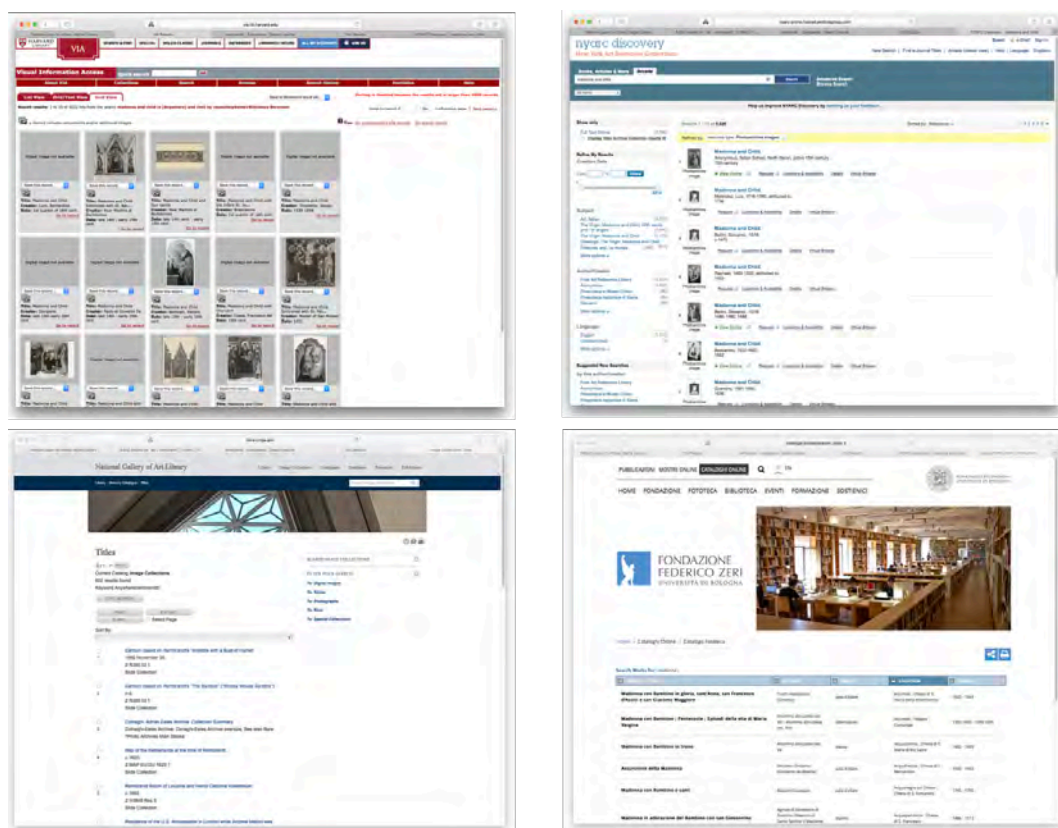


Figure 4 - Examples of catalogs in use for Photographic Archives

A testament to the lack of innovation in the field of image catalogs, is with the case of Harvard University. With digital collections containing many millions of objects and substantial financial resources to fund a new image catalog, the institution chose to move its legacy system to SharedShelf for the cataloging portion, and Ex-Libris' Primo catalog for the discovery interface. This decision was made in 2017 after five years of various committees exploring the options on the market. The previous system, VIA (Visual Information Access) was built in-house specifically to address art documentation needs. Its successor SharedShelf (which publishes data to ArtStor), is a union catalog that is shared by over 150 institutions across the United States. Being an existing system with a predefined data model, the system has drawn much criticism from collections managers who have witnessed meticulously cataloged object records, the result

of decades of work by institutional catalogers, get flattened out by a model that is lacking in flexibility expressivity. Additionally, as a subscription-based service that is closed-access, ArtStor has received a great deal of criticism, particularly from universities and institutions in Europe who are embracing open-access policies.

To-date, the Fondazione Zeri is perhaps the only collection of historical photographs that has managed to digitize and catalog its entire collection. A small part of this data has been made available online through an RDF data dump⁴⁵ and SPARQL endpoint, allowing for direct access to the dataset. Since the metadata in the public catalog is very rich, it also provides very good search functionality, and is perhaps the best example of a basic functional image catalog for historical photographs.

Information Retrieval and Discovery

Universities have by now become accustomed to the fact that the vast majority of scholarly research practice in search of images begin with Google. The search giant is undoubtedly able to provide the highest recall for the vast majority of searches. For users who are looking for quick access to an image to reference, Google will in most cases provide the path of least resistance. On the flipside, users aiming to perform in-depth research on a particular work of art that may be documented in many image repositories, face many more obstacles and challenges. Disparate documentation standards, languages, and most importantly poor indexing implementations make searching challenging and cumbersome. Image collection portals are

45. Daquino, Marilena, et al. *Zeri Photo Archive RDF Dataset*. Alma Mater Studiorum - Università di Bologna, 2016. *DataCite*, doi:[10.6092/unibo/amsacta/5157](https://doi.org/10.6092/unibo/amsacta/5157).

usually siloed repositories without any capacity for federation and most of them are not properly represented in Google search results. Although we do have portals for searching across collections of images, the main issue that has been brought forward is that in order to take metadata from many disparate collections, that data has to be flattened out and simplified so that it can fit into a prescribed container that is dictated by that platform. Paradoxically, the more images a platform contains, the more the data is “dumbed down” and loses its richness.

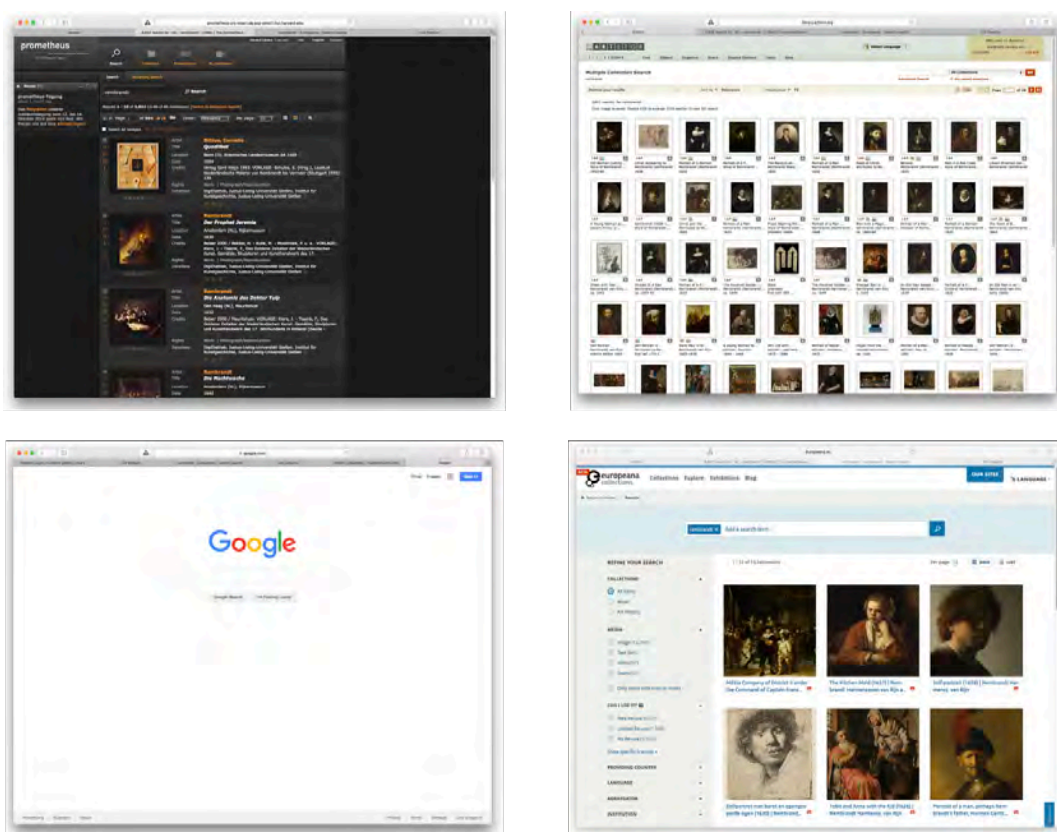


Figure 5: Image portals

Europeana is one of the few open-access portals that has tried to tackle the challenge of federation, but still struggles with issues of precision and relevance ranking, since the system is

essentially a full text index of records with little to no structure. Despite the underlying data model accounting for fields like “title” and “creator”, there is no way to search specifically within these fields. Relevance ranking is also problematic, with results shifting dramatically over time. A search for “Mona Lisa” in August of 2018 displayed a pair of shoes and a handbag as the first result. Six months later, the first result was an image of a pizzeria named “Mona Lisa”. Although depictions of the Mona Lisa are also returned as part of the results, there is no way to filter these results in a systematic fashion, making these tools much less useful to scholars with more complex research questions.

Subscription-based portals like ArtStor and Prometheus do offer better precision, but are limited in scope and do not provide metadata beyond the basics required to find the image. They are systems whose primary function was to provide access to collections of images that Art History faculty can use to project in their lectures. They are not systems built to document or retrieve data about the complex network of knowledge surrounding works of art, or enable any kind of serendipitous discovery that allows for the discovery of related works. Since these systems operate on relational databases with flat data models, data context is not queryable, and therefore is problematic for both information retrieval and discovery.

Research Systems

Thus far, digital platforms for publishing data that support the intellectual inquiries of scholars have largely been created ad-hoc for a limited set of use-cases, often resulting in a cumbersome apparatus, limited in scope and functionality, often requiring advanced knowledge

of programming languages. Common tools such as Omeka⁴⁶ for digital publishing and ArcGIS⁴⁷ for digital mapping are limited to data input and publishing, and they lack the functionality required to drive more interpretive components of scholarship that move beyond full-text search. Tools that, in contrast do provide interpretive functionality (such as Gephy⁴⁸ and R⁴⁹) require large data sets and the assistance of computer scientists to prepare, parse, and curate the data. These tools place a burden on IT departments, and often require extensive training in computer programming and data management. To date, we have not seen a web-based publishing platform that supports the advanced research needs of scholars looking to perform in-depth analytical searches, let alone annotate or contribute knowledge back to platform.

Perhaps the only platform that focuses on collections and research data, placing emphasis on knowledge building and the research lifecycle, is ResearchSpace⁵⁰, a collaborative research environment in development for the past ten years at the British Museum. The project uses the Metaphactory middleware⁵¹ at its core, and builds components on top that can power the full lifecycle of digital scholarly research. The open-source collaborative Semantic Web environment is designed to use and build knowledge about the world and its history, and for this reason it was chosen to serve as a starting point for publishing the collections of the historical photo archive at the Harvard Center. Originally built to publish the collections of the British Museum, the platform can serve as both a documentation system and as a research platform. Its feature set is

46. Omeka. <https://omeka.org/>. Accessed 23 Feb. 2019.

47. ArcGIS Online | Interactive Maps Connecting People, Locations & Data. <https://www.arcgis.com/index.html>. Accessed 23 Feb. 2019.

48. Gephi - The Open Graph Viz Platform. <https://gephi.org/>. Accessed 23 Feb. 2019.

49. R: The R Project for Statistical Computing. <https://www.r-project.org/>. Accessed 23 Feb. 2019.

50. ResearchSpace - a Digital Wunderkammer for the Cultural Heritage Knowledge Graph. <https://www.researchspace.org/>. Accessed 10 Sept. 2017.

51. Metaphactory. <https://metaphacts.com/product>. Accessed 23 Feb. 2019.

rich and robust: storage, querying, inferencing, search, visualization, and authoring, all using well-established and open standards of the semantic web that enable the reuse and repurposing in Linked Data environments. As an open-source product, the platform allows for enhancements from the community, which in turn can be contributed back to the source code. The software architecture is data-centric, allowing user interfaces to be built around any kind of research and collections data. In 2017, using ResearchSpace as a foundation, the Harvard Center launched a Samuel H. Kress Foundation-sponsored project⁵² that published a scholarly digital edition of Bernard Berenson's (1865-1959) *The Drawings of the Florentine Painters*.⁵³ In order to accommodate some of the more complex digital research projects the Center intends to undertake, a series of necessary customizations to the platform have been identified and are underway. Alongside existing functionality that enables the publishing of semantically enriched collections data using the CIDOC-CRM, enhancements to allow for the digital publishing of scholarly articles and historical archival documents with semantic annotations will allow for a seamless transition between text-based and graph-based data. Geospatial mapping components will allow scholars to annotate maps, both historical and contemporary, creating structured data from these annotations that will allow for subsequent programmatic analysis. The integration of Computer Vision and Machine Learning services (such as Google and Clarifai) will allow scholars to perform searches on artworks to find visual similarity between them, as well as build models to identify visual themes across vast collections of images, enabling a new kind of visual research that can scale to unprecedented levels. This same computer vision services will enable

52. Klic, Lukas, et al. "Florentine Renaissance Drawings: A Linked Catalog for the Semantic Web." *Art Documentation: Journal of the Art Libraries Society of North America*, vol. 37, no. 1, Mar. 2018, pp. 33–43. www-journals-uchicago-edu.ezp-prod1.hul.harvard.edu (Atypon), doi:[10.1086/697276](https://doi.org/10.1086/697276).

53. *The Drawings of the Florentine Painters*. <http://florentinedrawings.itatti.harvard.edu/>. Accessed 23 Feb. 2019.

scholars to transcribe words written by the same hand, in batch, across entire collections of archival documents, greatly facilitating some of the more cumbersome archival work that has been necessary until now. Additionally, Natural Language Processing services will eventually provide automatic entity extraction and encoding for persons, places, things, and events. These enhancements to the ResearchSpace platform are all driven by digital projects, each with a clear vision and specific research questions that grapple with broad, multidisciplinary issues.

Impact

Transitioning important historical collections online into research environments that allow for the full expressiveness of complex research data, is work that is long overdue in the field of digital heritage. There is a real need for hybrid systems that couple documentation and research, as well as tools and methodologies that allow for the publishing of material in efficient ways that empower and embolden institutions to take on large projects that move their print collections into the digital sphere. With advancements in imaging technology, the digital capture of documents is no longer the greatest bottleneck for digitization projects. On the other hand, the time required to catalog documents is increasingly scarce as human capital within heritage institutions becomes increasingly scarce. Therefore a shift in institutional strategy is necessary, as it has become evident that fully cataloging all of the printed documents contained in our archives is an unsurmountable task. The approach taken in this project is to provide initial access to the material through a core set of metadata that can enable information retrieval. Researchers

may then augment the metadata of these collections in order to make them more accessible.

Linked Open Data, Computer Vision, Machine Learning, and Natural Language Processing are all methodologies that can facilitate and enhance access to this material without the need for a substantive intervention from collection staff.

This project broadly aims to share and disseminate a series of methodologies and digital tools that can serve other institutions in their quest for publishing rich historical collections as Linked Open Data in research environments that enable complex queries that move beyond full-text indexing. Linking archival collections to other datasets, including the Getty ULAN & AAT, GeoNames, VIAF, and Wikidata and others will contribute to the flourishing LOD cloud, in turn fueling greater interest in the field and facilitating widespread adoption. Although this publishing paradigm shift is a slow one, it is necessary to ensure steady growth in the humanities and sciences. Expanding Villa I Tatti's collections portal and providing a unified digital research environment will be of exceptional benefit to the field of early modern studies and the humanities more broadly. Scholars globally, would have free and open access to some of the highest quality scholarship in the field, generated at the Harvard Center. Current and former appointees, would be able to contribute research data back to the platform, in the form of micro publications and narratives that would be coupled with assertions about historical documents, actors, places, and cultural heritage objects. The research platform could also serve as a pedagogical instrument that could be incorporated into the curriculum of courses across the world, both in-person and in Massive Open Online Course (MOOC) environments. The nature of this machine-readable data lends itself well to a more playful and serendipitous discovery, making it attractive and engaging to both undergraduates and seasoned scholars alike. The

Harvard Center would ensure the exceptional quality of the scholarship, by mediating a peer-review process through which data gets published. Since these methodologies will be openly shared, they can serve to guide other institutions in their digital publishing efforts. Finally, this semantically enriched and structured data, would in turn contribute to a vibrant culture of open scholarship and collaboration among researchers, disrupting barriers posed by proprietary databases where information is kept in silos.

III - Capturing & Cleaning Collection Data

In building a path and set of best practices for moving from printed archival material to a digital environment, one must begin with an assessment of the collection. No two archival collections are alike, and different materials require shifts and adjustments in methodologies. Although the processes outlined in the forthcoming pages is specific to one collection, its methodology can be transferred to many other institutional collections of printed material, in particular those containing a minimal inventories or indexes that can serve as a starting point for processing the collection. The granularity, structure, and cleanliness of the data in the source index, will greatly impact the efficiency and usefulness of the resulting published material. Once having performed this initial survey of the source material, analyzing it's ability to be digitized and transformed, the methodologies and workflows for the entire project should be mapped out, as unforeseen obstacles can easily create roadblocks that either set the project back or make it impractical to move forward. A user-centered focus and a clear understanding of the desired result is instrumental in order assess the value of each step along the way, as obstacles in the transformation of the source material can cause roadblocks, some of which may not be worthwhile to address. While it is not possible to anticipate all of the potential challenges for such projects, if multiple stakeholders are involved, it is crucial that a single individual has intimate knowledge of all aspects, ranging from the content of the data to the way the software is implemented.

While it is possible to digitize collections and make them available without metadata, toolsets that can automatically process these collections are not mature enough to produce meaningful results. The greatest obstacle to creating workflows that are efficient and fast-paced is keying in data. If project stakeholders need to manually enter data, even if just a simple transcription, this process has the potential to take more time than all of the other tasks combined. Whenever possible, it is important to build workflows where users can select from a list, or use an auto-complete function. Limiting collection processing to use either the keyboard or mouse is also preferable, minimizing the amount of time spent on mechanical movements.

The Berenson Archive

The collection at Villa I Tatti was originally founded by Bernard Berenson, a prominent art historian who later donated his home to Harvard University for use as a research center. The center houses numerous collections of cultural heritage, including his original art collection, comprised of circa one hundred and fifty western and non-western works of art. The historic photograph archive holds over two hundred and sixty thousand photographic prints⁵⁴, many of which are long since damaged, destroyed or lost works of art and of which they are the only extant record. The result both of purchases and of important donations and bequests, this still-growing collection contains photographs of artworks in many media ranging from Antiquity to the middle of the 20th century, over eleven thousand of which have no known current location and are presumed to be lost or held in private collections. Finally, the historical archive, beginning with

54. *Photograph Archives | I Tatti | The Harvard University Center for Italian Renaissance Studies*. <http://itatti.harvard.edu/berenson-library/collections/photograph-archives>. Accessed 24 Feb. 2019.

the extensive papers of Bernard and Mary Berenson, has grown through donations or acquisitions of other collections, including the archive of the Committee to Rescue Italian Art (CRIA⁵⁵), created after the 1966 Florentine flood and headquartered at the Harvard Center. Providing online access to this material has and will continue to serve as a testing ground for publishing additional scholarly material and cultural heritage.

The historical photo archive originally served as Berenson's study collection, and was housed below the books that referenced a particular artist. Following Berenson's death in 1959, the photo archive has seen numerous transformations, including being moved to a dedicated building. Like many photographic archives, the photos are stored in boxes which have seen numerous reorganizations over the years. At present, they are organized according to school, artist, medium, and finally topographically according to location.

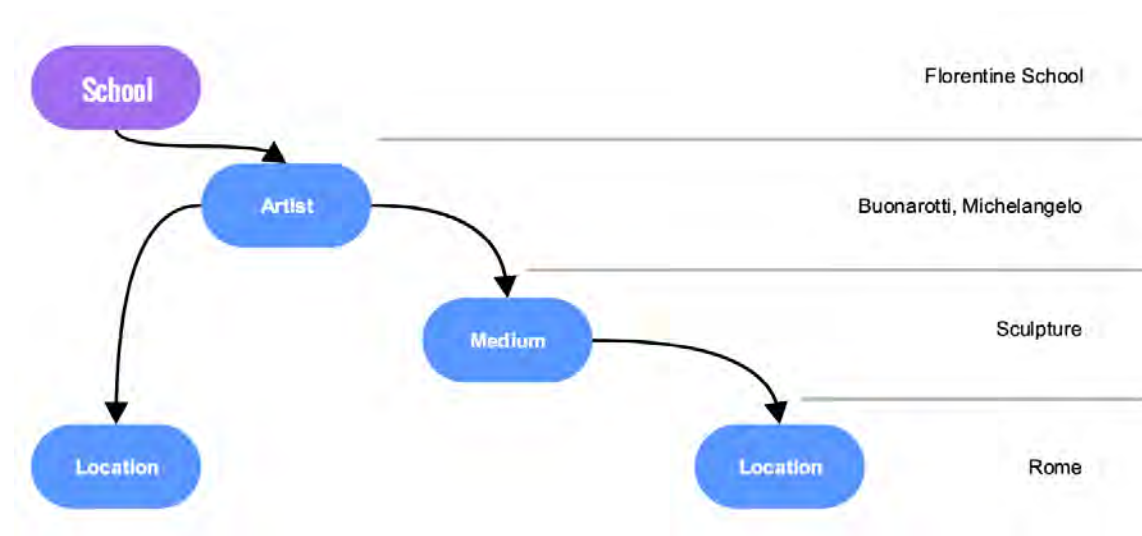


Figure 6 : Physical organization

55. CRIA - Committee to Rescue Italian Art. <https://cria.itatti.harvard.edu/>. Accessed 24 Feb. 2019.

Through a regular process of reattribution, the photographs are regularly shifted from one box to another. They are not mounted on any kind of backing, have no precise order within the box, and they have no identifier that will allow us to identify them as an individual object, aside from a small subset of 17,000 photographs that are classified as “homeless”. The lack of an identifier for each object was perhaps one of the greatest obstacles when trying to publish these materials online: without the ability to link physical objects to their digital surrogates makes the publishing process very challenging.

Four existing inventories of the collection, compiled at different times for different objectives, aim to document the collection at differing levels:

- SharedShelf: fully cataloged images from the Homeless Painting of the Italian Renaissance project (circa 11k records)⁵⁶
- Mappatura: a Microsoft Access database containing a folder-level inventory of the collection
- Collection-level records at the artist level providing a broad description of the photographs for each artist, stored in Harvard’s Integrated Library System⁵⁷ as bibliographic records in MARC format.
- Collection-level records by groups of 10–50 images, inventoried by catalogers from the Getty Research Institute between 1980–90.

Each of these inventories has served a different function at the time of creation, and there have not been any efforts to systematically bring these inventories together as they do not align with one another in a way that would lend themselves well to any integration process. The

56. *Homeless Paintings of the Italian Renaissance* | *I Tatti* | *The Harvard University Center for Italian Renaissance Studies*. <http://itatti.harvard.edu/berenson-library/collections/photograph-archives/homeless-paintings>. Accessed 24 Feb. 2019.

57. *HOLLIS*. https://images.hollis.harvard.edu/primo-explore/search?vid=HVD_IMAGES&sortby=rank&lang=en_US. Accessed 24 Feb. 2019.

metadata formats are all different, with custom data models that make common integration methods impossible without substantial transformation. The only possible alignment between these records would be at the photograph level, where each record would inherit the corresponding collection, box, artist, and work metadata as a property.

Previous Collection Publishing Efforts

The desire to systematically publish the entire collection of the Harvard Center was in many ways grounded and informed by past experience. Between 2007 and 2013, an Andrew W. Mellon and Samuel H. Kress Foundation-sponsored project was undertaken to digitize and catalog a collection of photographs of particular historical significance that were designated as “homeless” by Berenson, meaning that they were lost either to private collections, wars or simply their location was unknown.

The project, which employed roughly six full-time staff over a period of six years, resulted in the cataloging of less than ten percent of the collection (17k photographs representing 11k works of art). Continuing along that pace, it would take another eighty-five years to be able to digitize and inventory the remainder of the collection. In comparison to today’s standards, the methodology employed for that project could be perceived as rather rudimentary however: a single operator utilized a flatbed scanner to scan the photographs and upload them to the university digital asset management system, while a team of part-time catalogers conducted research on each object of art and created corresponding records in the collections management system. The process of digitization was particularly slow, with the flatbed scanner taking up to

five minutes to scan each side of the photograph. Subsequent post-processing included a completely manual cropping and rotation process, file naming according to the inventory number, and a series of other time-consuming tasks that resulted in an extremely high cost-per-image ratio. Since the artworks were presumably lost and completely lacking in documentation, the cataloging process was also particularly onerous, requiring extensive research for each object, attempting to locate it in a collection and performing the archival excavatory work necessary to make the record useful to scholars. The project overall, required substantial financial commitment, making it clear that in order to publish the rest of the collection, alternative solutions would need to be crafted. Expectations for the quantity and quality of metadata were going to need to be curbed, along with those of the images.

Getty Photographic Campaign

In the 1980's, the research arm of the Getty Center in Los Angeles (today renamed the Getty Research Institute) funded a series of projects in important historical photographic archives around the world, one of which was at Villa I Tatti. The objective was to make copies of the photographic prints contained in the collection, as preventative measures for preservation reasons (in case of fire or flood) and also to make the content available to scholars on opposite sides of the globe. A photographer was funded for nearly ten years to systematically photograph the entire collection. Due to the high cost per image, conditions were set to skip any duplicates or images of poor quality, especially if there were no annotations on the verso that were of any interest. The photographer used large, uncut reels of 35mm polyester B&W film, and a custom-

built camera mounted on a copy stand. The photographer experimented in the first reels testing the quality of the film and sequence of images, but he later settled on a workflow where the recto of the photograph was shot on a high-quality B&W photographic film, and the verso was shot on lower-quality microfilm as it only contained handwritten notes. This workflow was established in order to save on the cost of the project, as well as to streamline the imaging process, allowing the photographer to use two copy stands, shooting the verso of a photograph on one and the recto on the other. Each reel of photographs contained roughly five hundred shots, either of a recto or verso. Once a reel was finished, the photographer created prints which he sent back to California and occasionally other collections around the world. At the Getty, the photos were inventoried and some minimal records were created at the artist level. The project resulted in 455 reels of film containing a total 230,000 images, representing 130,000 photographs (one for the recto and one for the verso). This photographic documentation represented the core nucleus of the collection as it was in the 1980's.

The photographer created targets that were used to break collections of artworks into smaller chunks, possibly based on folders that were within the boxes at the time. Based on these groups of images, minimal collection-level records were created by Getty staff.

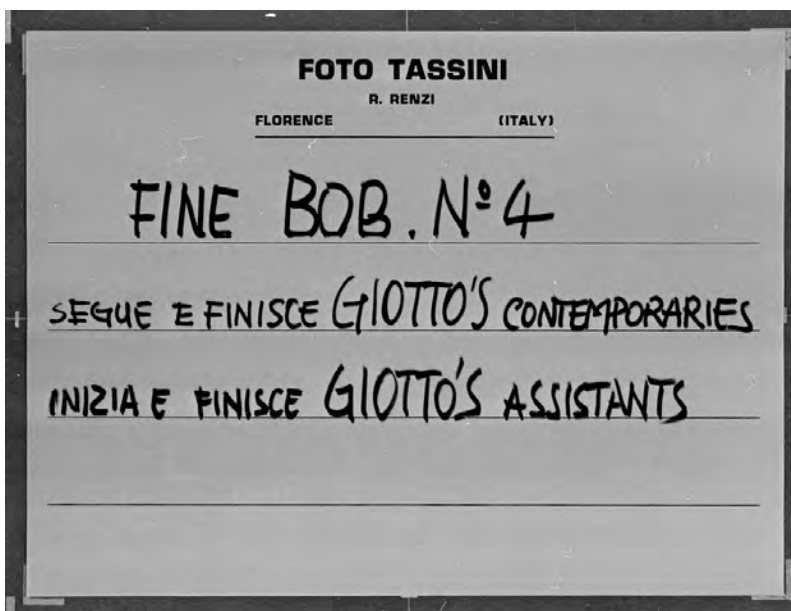


Figure 7: targets used to delineate groups of photographs

Since the inventory created at the Getty did not align with any other inventory or physical layout of the collection, it was not possible to align these data with other inventories such as the “Mappatura”, the database that documents the physical organization of the collection.

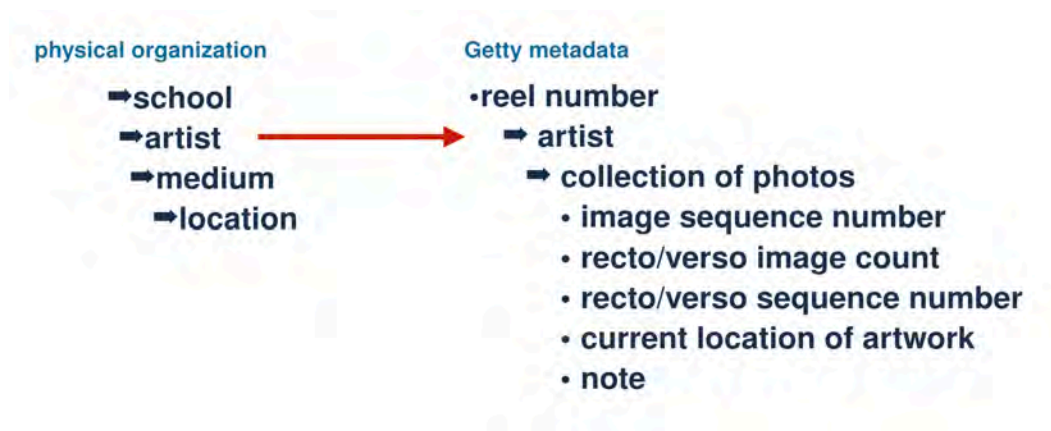


Figure 8: alignment of Getty inventory to current physical organization

The lack of alignment meant that in order to be able to use these metadata, new records would need to be created at the photograph level, linking to metadata from the collection records. The first portion of the Getty collection record contained administrative metadata, including the number of the film reel, the number of photographs in this group, as well as their position in the reel, a key piece of data for being able to link up images.

```

42280 <Object_locs>Parma: Galleria Nazionale</Object_locs>
42281 <Object_locs>Pianore di Camaiore: Villa (stolen)</Object_locs>
42282 <Object_locs>Prague: Narodni Galerie</Object_locs>
42283 <Object_locs>Philadelphia: Johnson Collection</Object_locs>
42284 <Object_locs>New York: Myron Taylor Collection</Object_locs>
42285 <Object_locs>Paris: Louvre</Object_locs>
42286 <Object_locs>Parma: Pinacoteca</Object_locs>
42287 <Object_locs>Pesaro: Villa Imperiale</Object_locs>
42288 <Object_locs>Providence: Rhode Island School of Design</Object_locs>
42289 <Object_Notes>n. 305 not printed</Object_Notes>
42290 </record>
42291 <record>
42292 <recno>3230</recno>
42293 <Project_neg_>182 (354-388)</Project_neg_>
42294 <No._of_images>35</No._of_images>
42295 <COUNT>35</COUNT>
42296 <I_Tatti_ref_>XXV.20s</I_Tatti_ref_>
42297 <Artists_names>Dossi, Dosso</Artists_names>
42298 <Object_locs>Rome: Castel S. Angelo; Campidoglio; Galleria Nazionale Barberini; Galleria Borghese;
42299 <Object_locs>Vatican City: Pinacoteca Vaticana</Object_locs>
42300 </record>
42301 <record>
42302 <recno>3231</recno>
42303 <Project_neg_>182 (389-410)</Project_neg_>
42304 <No._of_images>22</No._of_images>
42305 <COUNT>22</COUNT>
42306 <I_Tatti_ref_>XXV.20s</I_Tatti_ref_>
42307 <Artists_names>Dossi, Dosso</Artists_names>
42308 <Object_locs>Venice: Cini Collection</Object_locs>
42309 <Object_locs>Vienna: Gemäldegalerie; Lanckoronksi Collection; ex Baron Tucher Collection</Object_lo
42310 <Object_locs>Washington, D.C.: National Gallery of Art</Object_locs>
42311 <Object_locs>Wichita, Kansas: Art Association</Object_locs>
42312 <Object_locs>Worcester, Mass.: Art Museum</Object_locs>
42313 <Object_Notes>nos. 390, 404, and 407 not printed</Object_Notes>
42314 </record>
42315 <record>
42316 <recno>3232</recno>

```

Figure 9: Sample of Getty metadata

The next portion of the record, contained the artist name and a list of locations where the artworks were held, according to the handwritten text on the back of the photograph. Additionally, a notes field provided either additional administrative metadata (such as “305 not

printed” as seen in Figure 9), or notes pertaining to the content of the group of images (such as “drawings”). These notes were not structured in any way, so they were not usable beyond being encoded as a simple note field. The Artist name field, although not normalized across the entire group of records, was clean enough to be able to de-duplicate and reconcile. The location field, described as “Object_loc”, was structured as a city followed by a colon “:” and a list of institutions separated by a semicolon “;”. This semi-structured data lent itself very well to cleanup and parsing, allowing a simple relational database to be built. The value of this provenance information was deemed as being very useful for the Art History community, as this information may be the only place where it is documented.

FotoIndex Publishing Strategy

Given the extent and complexity of the collection, a feasible strategy for publishing the contents of the archive was an essential first step in moving forward. Scanning the original photographs and generating new metadata for them, even with the latest advancements in imaging technology was not within the financial means of the institute. The films reels were by far the most attractive option for digitizing the collection as they could be scanned in batch with very little manual intervention, and the collection-level records would provide a solid subset of semi-structured metadata that could be associated with each image. It was here that the “FotoIndex” project was born, with the objective to digitize these reels and create a base index that could then be enriched later. Although the quality of a scan of a photograph of a photograph

is clearly not equatable to scanning the original, they could serve as reference copies for users and they would be searchable by artist, institution, as well as through visual search means, using other images. If the user wanted a higher quality image this could be requested through the archive. Most importantly, the project would provide access to the verso of the photographs, which contained over a century of annotations by trusted scholars.

Following a digitization of the reels, verso and recto images would need to be linked to one another, and this image pair would need to be connected to the collection-level metadata. Since the collection-level record contained a list of locations where all of the artworks within that group were held, individual locations would need to be linked to image pairs. This process would need to be implemented in such a way that a pair of catalogers could pass through all 230,000 images in a single pass, performing all of the necessary linking with a minimum number of clicks. The creation of records at the item-level would also provide an opportunity to integrate all four collections of metadata that describe the photographs.

Scan

Given the decision to digitize the reels of film, certain considerations were made regarding the in-sourcing/outsourcing of this process. The archive did not own other collections of 35mm film, so the option of purchasing a film scanner with batch functionality explicitly for this process would have to outweigh the cost of outsourcing. Although professional scanners that can handle 35mm film in batch (such as the Nikon Coolscan 5000 scanner⁵⁸) were relatively

58. *Super COOLSCAN 5000 ED from Nikon*. <https://www.nikonusa.com/en/nikon-products/product-archive/film-scanners/super-coolscan-5000-ed.html>. Accessed 24 Feb. 2019.

economical, they were all limited to films that have perforations running along the edge, and a maximum of forty frames, the maximum number of frames within a commercial roll of film. Since the reels had over five hundred frames, the only option was a microfilm scanner that had a sensor of sufficient quality similar to a film scanner. The cost of this unit in comparison to outsourcing the project to a vendor — which included initial cropping for the frames, image rotation, and quality control — was more than double the cost.

It was therefore decided to outsource this portion of the project to a vendor. The batch scanning of the reel itself proved to be a rather quick process. The vendor delivered TIFF files though Google Drive as the reels were digitized and the initial post-processing was performed.

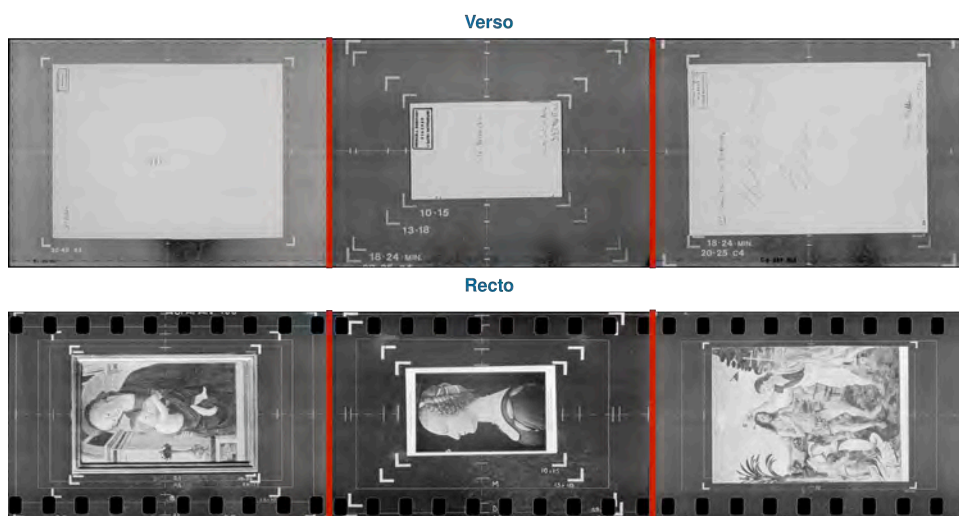


Figure 10: sample files returned from the vendor (red line indicates crop)

As illustrated in figure 10, initial frame detection cropped the images in a way that separated one from the other, but left the rest of the frame with perforated edges and the copy board background. Since the copyboard portion of the image was of no interest to users, it was

necessary to crop down to the photograph and create a derivative, still preserving the original to provide context, such as the physical dimension of the object within the frame.

Crop

The necessity to crop the image programmatically was very clear, not just for reasons of visual aesthetics but also to make the image more searchable with visual similarity engines (the background would prove to be quite distracting for most visual search algorithms).

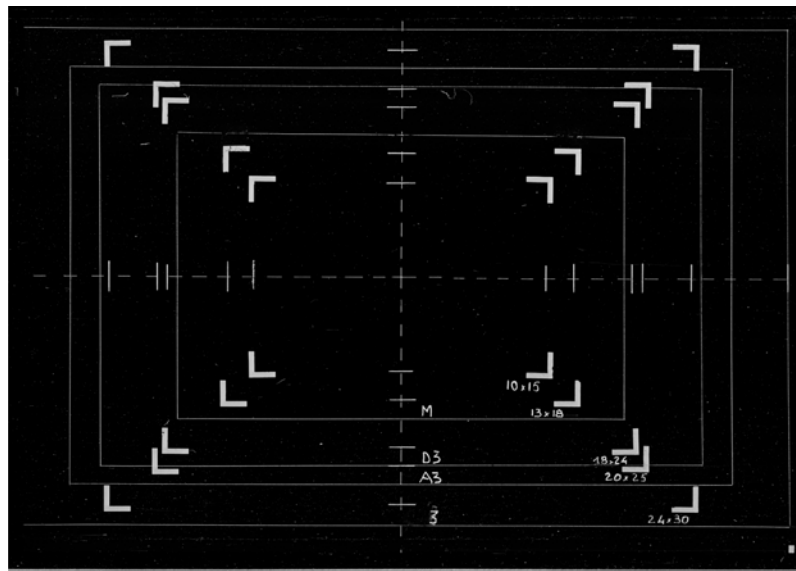


Figure 11: Sample copyboard background

Numerous initial attempts were made using readily available solutions off the shelf, including Adobe Photoshop auto crop and straighten functionality, as well as various combinations of ImageMagick⁵⁹ command-line functionality. As seen from the sample in figure 12, Photoshop was generally unable to differentiate the copy board from the photograph, even with some initial pre-processing on the contrast. Additionally, the lack of command-line tools meant that reproducing the crop on master images would have been impossible, as the program was not able to return the image coordinates of the crop. Although the Photoshop auto-crop functionality outperforms many other products on the market, it is more geared towards images that have been scanned in a controlled environment, where the background of the image can be more neutral.



Figure 12: Photoshop auto crop and straighten functionality

59. LLC, ImageMagick Studio. "ImageMagick." *ImageMagick*, <https://imagemagick.org/>. Accessed 30 Jan 2019.

Experimentations with ImageMagick proved to have similar results. Testing image pre-processing with functions such as “fuzz”, “trim”, and “repage” proved similarly problematic. The main issue being that that ImageMagick would always interpret the absolute black background where the film perforations started as the edge of the photograph. Figure 13 shows a sample of this result, where the crop is brought down to the border of those perforations, even with a “fuzz” parameter set at 99%. These results varied with images of differing contrast, but generally proved to be unreliable.

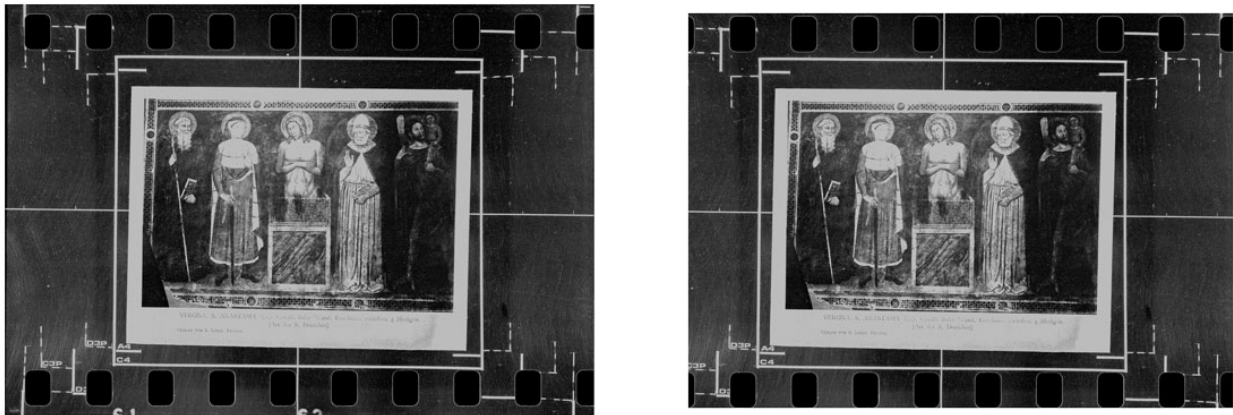


Figure 13: ImageMagick crop result sample

Although both Photoshop and ImageMagick did provide occasionally useful results, these were limited to small subset of images that had a much clearer delineation between the photograph and the copyboard background, and when there were few or no measurement markers visible on the copyboard background. This lack of stable reproducibility of successful results meant that these tools were not a good fit the cropping task.

Additional tests were performed using background extraction methods, initially attempting to use a sample copyboard with no photograph as seen in Figure 11. This attempt

proved to be unsuccessful due to substantial changes in the background as the photographer zoomed in and out on each image. Other tests, based on the methodologies published by A. Rahman⁶⁰, using a GrabCut algorithm that “provides a way to get a segmentation of a target object with minimal input from a user and extracts it as foreground” were also tested. The sample code, based on this attempt is listed in Appendix A. The results produced from these tests were far more useful as they result in the copyboard background being replaced by an absolute black, which could be subsequently be cropped down. A sample result of this test, as illustrated in Figure 14 brought forward several issues with the images themselves, where it became evident that the angle at which they were shot, together with lens distortion, resulted in images that were non-rectangular.



Figure 14: GrabCut methodology sample

60. Rahman, Abrar. *Interactive Foreground Extraction with Superpixels*. p. 54.

The results were however unstable due to contrast issues, either with certain portions of the images or with entire groups of images. Given the necessity to follow cropping standards set forth by the US-based Federal Agencies Digital Guidelines Initiative (FADGI)⁶¹, along with the Library of Congress (LOC) *Technical Standards for Digital Conversion of Text and Graphic Materials*⁶², the GrabCut methodology in its current form would not comply as it creates an image crop that would remove the borders of the photograph in most cases. These standards require the “presentation of the entire original sheet or page. In no event shall the actual document be cropped”. Although some feathering or adjustment to the margins of the crop could have been implemented, this was overshadowed by the fact that roughly 30-40% of the images had such a low contrast with the copyboard that it would have required a manual intervention on a substantial portion of the images.

In most scenarios where material is being digitized in a controlled environment, the issue of cropping is usually trivial, as various methodologies (such as background extraction for a particular color) can usually achieve near perfect results. In the case of scanning the film from the FotoIndex project, the fact that it was shot under various lighting conditions, finding a one-size-fits-all solution for cropping was far from trivial. Following a long series of subsequent experiments, a script to find the coordinates of the image borders using OpenCV was implemented.

61. *Federal Agencies Digital Guidelines Initiative*. <http://www.digitizationguidelines.gov/>. Accessed 2 Mar. 2019.

62. *The Library of Congress Technical Standards for Digital Conversion of Text and Graphic Materials*. p. 15.

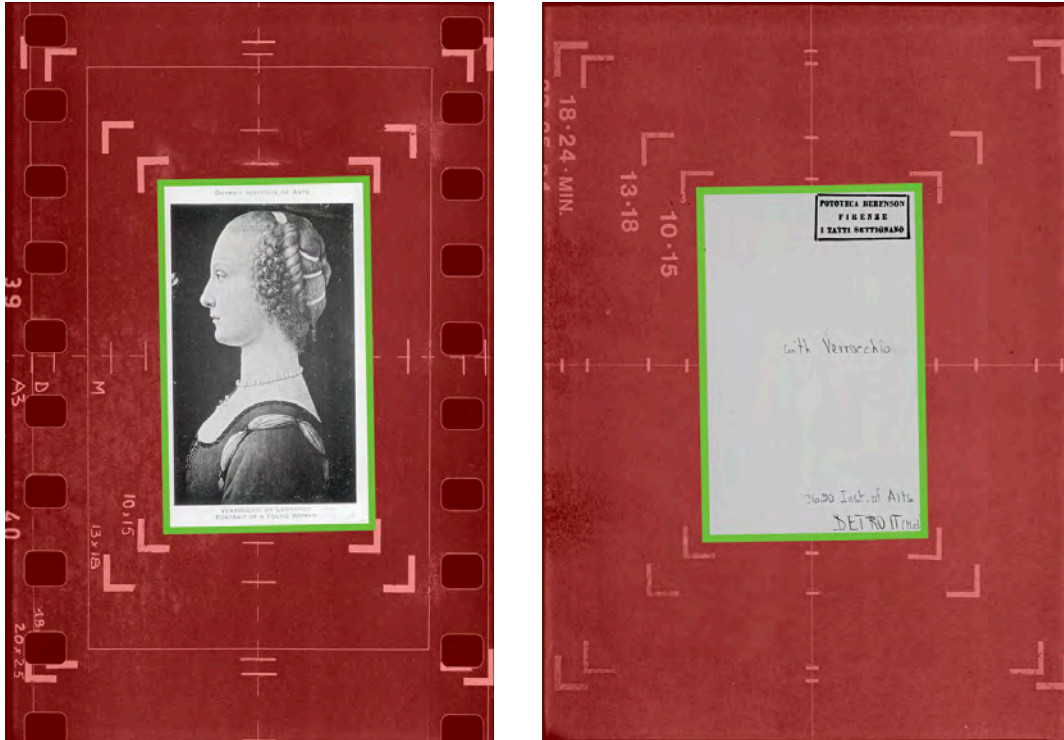


Figure 15: Desired crop for an image

Beginning with some preprocessing of the images including denoising (fastNlMeansDenoising), improving the contrast (Contrast Limited Adaptive Histogram Equalization), image thresholding (Otsu's method) and applying a blur to the remaining part of the image, the findContours method was able to draw a bounding box accurately around the images. Given the disparate nature of the background contrast and photograph placement within the image throughout the reels, the final script had to be re-run multiple times on different subsets of images with different variables. The final script that was used to for this cropping process is available in Appendix B. Although they required a considerable amount of experimentation with image preprocessing, with some modifications they may be usable on other collections of film as well.



Figure 16: *findContours* OpenCV functions to find the bounding box around the photograph⁶³

The results from these scripts produced a CSV file which contained image coordinates to pinpoint the location of the photograph, as well as a rotation percentage. Converting this output into a format that could be interpreted by a server that delivers images compliant with the International Image Interoperability Framework (IIIF), allowed the original images to be uploaded without cropping. Leveraging the IIIF API, links to the images were then built to return the cropped region of the image.

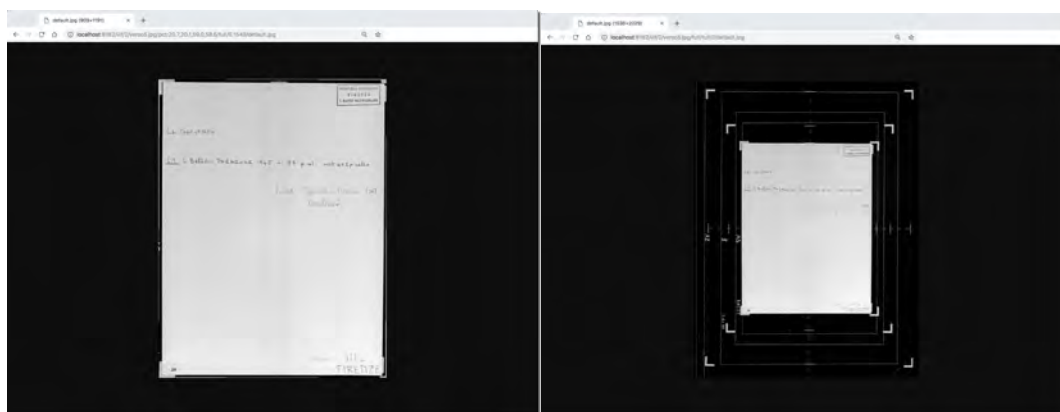


Figure 17: The same image being delivered with different views using IIIF image coordinates

63. Image credit: https://docs.opencv.org/3.0.0/d4/d73/tutorial_py_contours_begin.html

The first parameter of the URL, following the image file name, allows us to select a region in percentages as shown in Figure 18, as defined by the IIIF API specification⁶⁴.

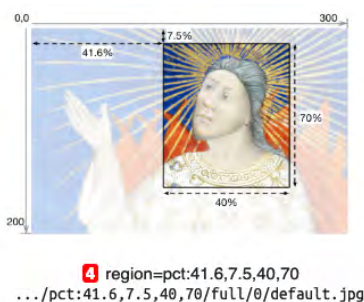


Figure 18: using IIIF image coordinates⁶⁵

The API is also able to account for image rotation, which was a common issue for many of the images, as the photograph was not perfectly aligned horizontally with the copyboard. The angle of rotation is the third parameter that can be passed into the URL, from 0 to 360 degrees. The versatility of the IIIF image API allows for the preservation of the original image that was produced by the scanner in TIFF format. Uploading those originals, the IIIF server is able to produce any kind of derivative on the fly, of any size that is requested. At the time these scripts were built (2016), DigiLib⁶⁶ was used as it was the most lightweight and versatile open-source IIIF server available. Since then, other products (such as Canteloupe⁶⁷) have emerged, allowing for a greater flexibility in terms of file management and caching, while still providing the same cropping and region selection functionality.

64. *Image API 2.1.1 — IIIF | International Image Interoperability Framework*. <https://iiif.io/api/image/2.1/#region>. Accessed 2 Mar. 2019.

65. *Image credit*: <https://iiif.io/api/image/2.1/#region>

66. *Digilib - The Digital Image Library* -. <http://digilib.sourceforge.net/index.html>. Accessed 2 Mar. 2019.

67. *Cantaloupe Image Server* . <https://medusa-project.github.io/cantaloupe/>. Accessed 2 Mar. 2019.

Most importantly, the IIIF framework allowed for the preservation of the original image. This in turn allowed for the crops to be reviewed by staff in batch with a cropping approval process where all images were copied to a IIIF server, and the image filenames and coordinates were loaded into a Google Doc spreadsheet. Combining the coordinates with the filenames, a URL for the image was constructed that was passed into the `image()` function of Google docs to visualize both the original image and resulting crop. For most reels, about 5-15 images (roughly 1-3% overall) needed to be adjusted manually, which could be done by simply changing the percent parameter that determined the four sizes of the crop. Given the ability to instantly visualize all images in a reel, and quickly change the crop parameters, this manual review process was performed in the span of a few days.

The three main edge cases that presented themselves during this phase of the project, were primarily images with very low contrast between the copyboard and the photograph, non-rectangular photographs, and the photograph extending beneath the perforations on the side of the film. Given the limited need to reproduce such cropping results (it is unusual that in the cultural heritage domain we are scanning film reels that are photographs of photographs), other methodologies that employ machine learning were explored but not implemented due to the amount of effort needed to implement. Nevertheless, the author acknowledges that in retrospect machine learning could have been leveraged for this process to create a more streamlined and reproducible workflow that would have been of greater interest to the digital heritage sector.

Measure

Since the copyboard background provided measurements to indicate the size, it was deemed that the physical size of the photograph was a piece of metadata that would be useful to provide to scholars, and would have been important in the future should a need arise to track down the original photograph. This task, which at the outset seemed very feasible, proved to be too arduous a task and not worth the time investment to complete properly. The challenges stemmed from the fact that at the time of shooting, the photographer zoomed in and out on the image plane to capture the photographs at differing focal lengths. Had he maintained a constant focal length, even for individual reels, measuring the size of the photograph would have been far more manageable.

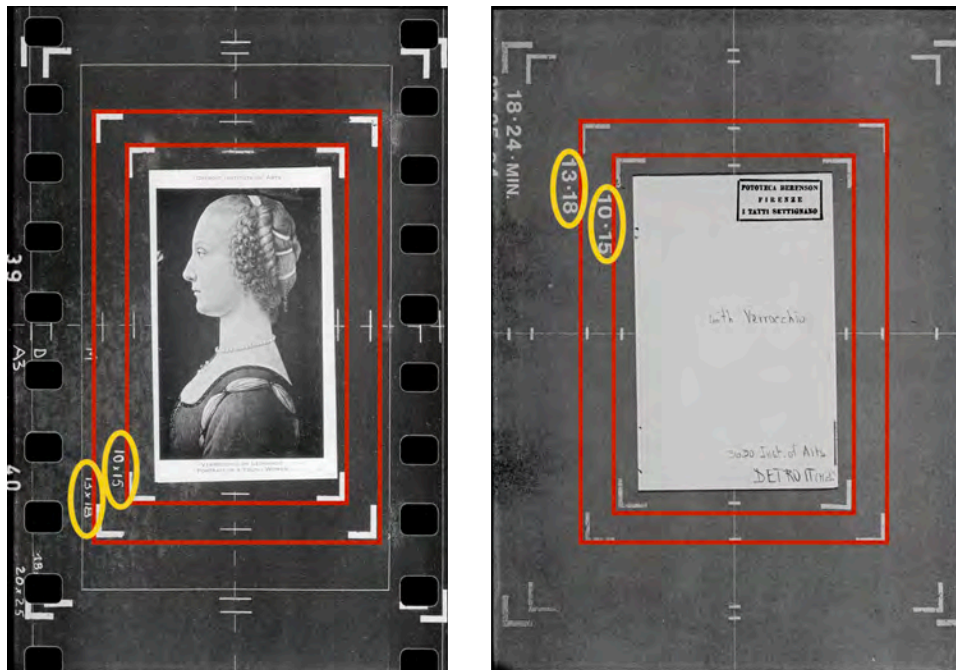


Figure 19: Measurement markers on the copyboard, allowing for the calculation of the physical dimensions of the photograph.

As is illustrated in figure 1, copyboard markers indicating the size could be used to calculate the height and width of the photograph. The first obstacle in this problem was to delineate a bounding box of the photograph, which was captured in the previous step during the cropping process, and calculate the difference between new bounding boxes that would need to be drawn for at least one of the markers on the copyboard. Here the number of edge cases made the problem far too cumbersome a task to be able to perform accurately at a large scale. Here, roughly 40% of the images were too large within the frame and did not reveal enough of the copyboard to be able to properly interpret the results programmatically. Most importantly, without a consistent mechanism to indicate problems with the measuring process, there was not way to systematically identify the useful results vs. those that were erroneous. It was also discovered, after measuring the actual photograph in the archive, that based on the focal length that was used to photograph the photograph, additional distortion was taking place on the aspect ration of the image. This distortion was not constant nor visible to the eye when inspecting the images, so coupling these factors with such a high rate of measurement failures led to ceasing further work on the problem. Instead, since an uncropped photograph was being provided to the user through the IIF viewer, the user would be able to infer the approximate measurements visually, should that be a piece of data that is of interest to them.

Collation

Subsequent to extracting semi-structured metadata from the original index provided by the Getty metadata, and processing all of the images for cropping, a data collation process involved breaking up all of the bits of data in a relational database, assigning identifiers, and deduplicating any entries, including artist names and institutions. The images that were provided by the vendor had file names that were structured according to their placement on the reel. Since the Getty index had administrative metadata that grouped collections of photographs specifically to particular positions on the reel, programmatically linking these records to the images was a fairly trivial task. However, anomalies in the enumeration of images within reels made it evident that they would need to be manually confirmed. These anomalies were caused by the fact that when the photographer made a mistake, he would reshoot the image, either on the recto or verso, resulting in duplicate images and a misalignment in the numbering of images on a reel.



Figure 20: an image that was discarded by the photographer

As seen in Figure 20, the photographer would usually punch a hole in the film when he chose to discard an image. At times the hole was not punched all the way through or there was no hole at all. This resulted in a need to pass through all of the images manually as it caused misalignment between the recto and verso images for all of the reels, as well as groups of images to the Getty collection records. The fact that the misalignment was only minor (less than 10 images per reel of 500 were discarded) meant that the alignment work could be completed in an efficient and light process that required subject experts simply to confirm the matching of a verso, recto, and catalog record.

Matchmaker

Following the data and image collation processes it became clear that in order to align images to the metadata, a custom application would have to be built so that catalogers can go through the entire collection. Since the application was only meant for internal use, and it would only be used once for this specific function, it was decided that Microsoft Access would be the most efficient tool with which it would be possible to quickly build this kind of tool. Although the application is a closed-sourced and any functionality would need to be built with MS Basic — a programming language that is somewhat lacking — working in a very controlled environment with a specific type of computer, monitor arrangement, and access to the large collection of image assets (both smaller jpg derivatives for loading quickly and full-sized images when one would need to zoom in), meant that the entire process could be completed very

quickly. While a web-based application would have been more desirable, the inability to have multiple application windows on different screens, manage image assets on institutional networked storage, and the amount of effort it takes to make changes to user interfaces, made it evident that MS Access would provide us with the required functionality to perform the task with the highest level of efficiency.



Figure 21: the Matchmaker application screen layout

As seen in figure 21, the application runs on dual-screen setup. The left screen is used to perform all of the tasks, while the right screen is used to put the image in context to the previous and subsequent images on the reel.

It was very important to maintain a balance of tasks throughout the process to ensure that the catalogers were not being overloaded, yet guarantee that they would not need to pass through the images more than once. Although it would have been possible to use the process to transcribe other content from the backs of the photographs, any transcription activities would have substantially increased the amount of time they would have had to spend on each image pair. As

a result, the following five tasks were identified as actionable with two clicks of the mouse or

keyboard shortcuts:

- a) Link image verso and recto
- b) Confirm alignment of Getty collection record to image pair (therefore linking an artist to an image pair)
- c) Apply provenance data to each image pair by selecting the institution.
- d) Discard poor quality images, or identify other issues
- e) Group images together that represent a single artwork
- f) Possibly correct attribution
- g) Select a primary, or display record for a work of art

These functions can be seen mapped out in figure 22 below:

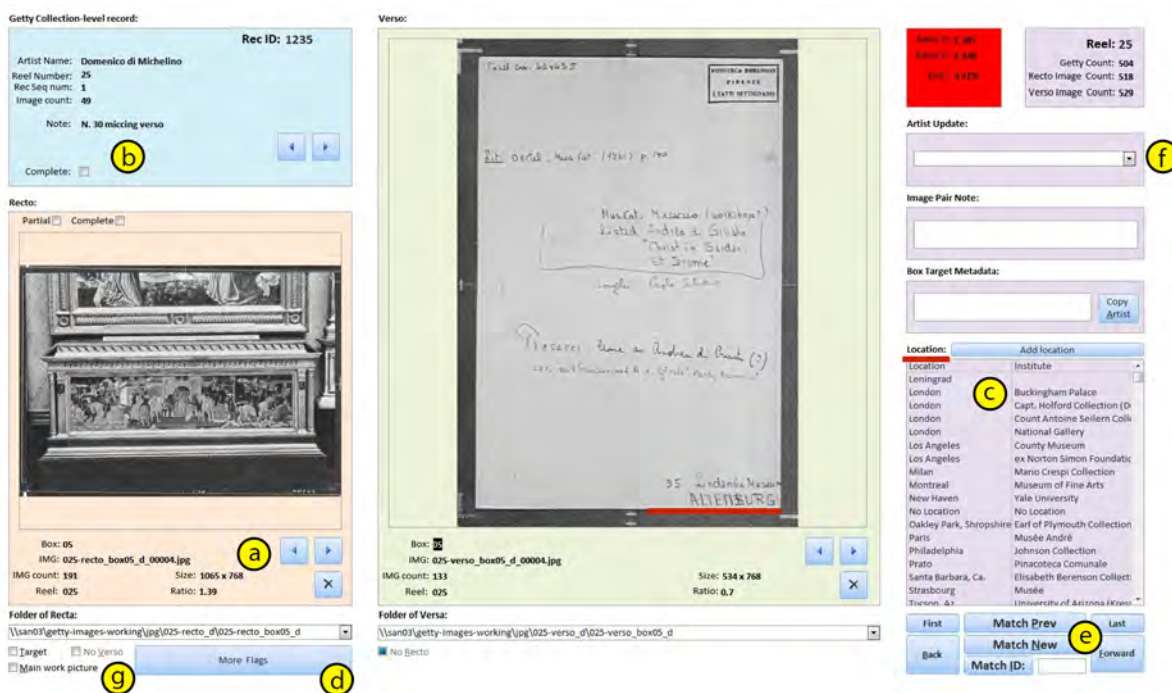


Figure 22: left screen of the Matchmaker application with corresponding functionality

The matching of the verso and recto image happened automatically as the user advanced through the pairs of images. The only time that their intervention was needed was when there was a misalignment in the pairs, in which case it would modify the alignment of all subsequent

images as well. In order to assist the cataloger in this process, the aspect ratio for each image was calculated. As seen in the top right of figure 22 with a background in red, the aspect ratio for the recto and verso images was calculated simply by dividing up the number of pixels for the height and width. This resulted in a number that was immune to changes in image orientation, since often the historical photograph may have been taken horizontally but the annotations on the back were written vertically. By dynamically calculating the difference between the aspect ratio of the two images, a threshold of three percent was set that would make the background of the ratio box turn red and let the user know it was unlikely that the two images were a correct match. This threshold was ultimately used as a guide, and was determined through some trial and error, since the aspect ratio was also affected by the crop of the image and other distortion resulting from the imaging process (either the photograph on the reel or the digitization process of that photograph).

The most frequent task for the user was to click twice for each image pair. The first click is to identify the location of where the artwork was held at the time, data which was handwritten on the back of the photograph, and also contained in the index data provided by the Getty, but not linked to a specific photograph. This was done by selecting from the list as indicated by the letter “C” in Figure 22. This list was extracted from the Getty index record, and is limited to the institutions that were listed for that particular record (usually not more than 10). Subsequent to selecting the location, the second click was either “Match Prev” or “Match New” as indicated by letter “E” in Figure 22. Since it was very common for there to be multiple images in a row that represented the same artwork, this allowed for the generation of identifiers for individual works of art, which would be connected to multiple pairs of images. “Match Prev” linked the photograph to the ID of the previous work, while “Match New” created an identifier for a new

artwork. Additionally, when one artwork had many images that include details, it was important to select the most appropriate representation of that work. Therefore, as the cataloger moved along, they would also be able to see the upcoming and previous images on the right screen. Based on being able to quickly look at the sequence of images, they would determine if a specific image should be chosen as a “display” record for that particular work of art. This is was a critical step necessary for displaying the artwork record alongside others in the user interface of an image catalog, since you generally show only one image to represent a particular artwork.

Finally, if the attribution was clearly incorrect, the cataloger would be able to select a new artist from a dropdown list. This was done in rare cases where the subject expert was immediately able to ascertain the incorrect attribution of an artwork without having to do any research. This functionality was added at a later date when these instances began to emerge throughout the matching process. Other functionality that had not been accounted for from the outset, was the need to add new institutions to the list that were not present. Fortunately these modifications did not impact the database structure for existing fields so the they were able to be added without disrupting the work that had already been done.

The author began the planning of the Matchmaker software by mapping out all of the required functionality based on the available metadata, then built user interface mockups in Gliffy. Subsequently the data was imported into various Microsoft access tables, linking up ID's for collection records, artists, images and institutions. The “forms” functionality of MS Access was used to build the user interface that allowed for the display of images, sub-tables to select data elements, and buttons that would trigger functions that wrote specific data to the various tables within the database. Similar to FileMaker Pro, MS Access provides a simple GUI for the

editing of these forms that made this process very speedy and efficient for rapidly developing applications that have a specific use-case. Once the application was built, one photograph cataloger performed sanity testing on a number of reels, and as issues were quickly brought to light they were iteratively adjusted at the application level. Following a few weeks of testing, it was determined that the application was ready for use in production and functionality was added to allow multiple users to write to the database at the same time with protection mechanisms to ensure their work did not overlap. After staff began using the software, weekly progress meetings were held to provide additional feedback on the software and to assess progress, making iterative adjustments along the way to streamline the process.

Matching Efficiency

Since task efficiency was a key factor in the matching process, functionality was added to ensure that the Matchmaker application also tracked the amount of time that it took to perform a single match. This was done by adding a username (taken from the name of the computer that was running the application) and a timestamp every time the “Match Previous” or “Match New” button was clicked. At the end of the matching process, the time between each match was calculated based on the username, discarding matches that took longer than 60 seconds, since it could be assumed that the staff performing the match would have taken an occasional break.

Initial objectives set a desired matching time of around five seconds per image pair, roughly the amount of time that was required for the eyes of the staff member to look at the recto and verso of the photograph, read the inscription, select the institution from a list, and click a

match button. For 115 thousand matches (115k recto images and 115k verso images) at five seconds per match, the collection could get processed in roughly 160 working hours. With two catalogers working part-time, it was originally estimated that the entire process could have been reasonably distributed over a period of three months.

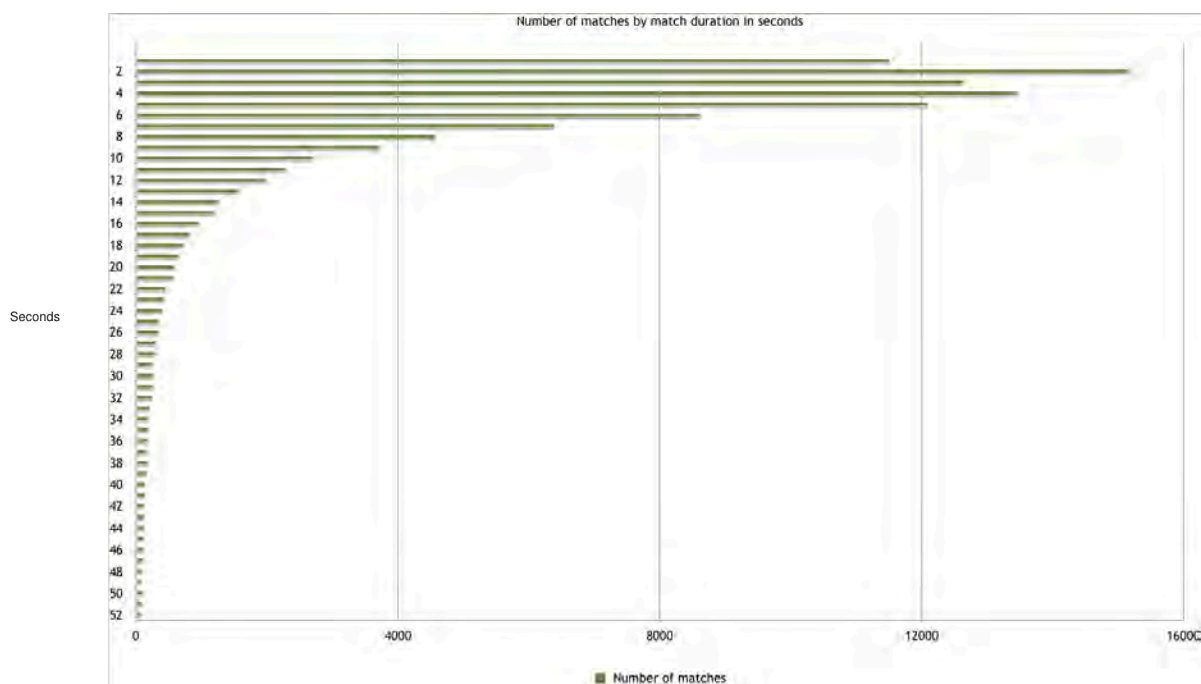


Figure 23: number of seconds per match

The actual times ended up being very close, with the entire process being completed in under four months. As illustrated by figure 23, the vast majority of the matches took between 2 and 8 seconds each (77%). The total time for the entire process ended up at 223 working hours, resulting in a difference of 73 hours. This miscalculation was the result of the remaining 23% of matches that took nine seconds or more, but due to one of the staff being able to dedicate more time to the project, the entire process was completed as scheduled.

Entity Reconciliation

Entity Reconciliation (or Entity Resolution) is the process of linking internal entities (persons, places, objects, etc.) to external sources to facilitate integration methods across the web of data. The process involves “identifying records that represent the same real world entity, and identifying records that are similar but do not represent the same real-world entity”⁶⁸. Linkages are then created between these identities by appending a common identifier to denote the fact that they are equivalent.⁶⁹ In the context of the Semantic Web, these links are generally created using the Web Ontology Language⁷⁰ (OWL) through a “sameAs” statement, or using the Simple Knowledge Organization System (SKOS⁷¹) statement “closeMatch”.

Aligning all of the key entities within a dataset is a critical component to be able to leverage many of the benefits of Linked Data technology. Defined as the fourth principle by Tim Burners-Lee in his 2006 article on Linked Data design issues, he states that one should “include links to other URIs, so that they can discover more things”.⁷² In the case of the FotoIndex project, this meant aligning terms with external vocabularies and reference sources commonly used in the Cultural Heritage domain. For artist names and artistic terms and techniques, the most comprehensive vocabularies are the Union List of Artist Names (ULAN) and the Art and Architecture Thesaurus (AAT) maintained by the Getty Research Institute. For institution names,

68. Enríquez, J. G., et al. “Entity Reconciliation in Big Data Sources: A Systematic Mapping Study.” *Expert Systems with Applications*, vol. 80, Sept. 2017, pp. 14–27. *ScienceDirect*, doi:[10.1016/j.eswa.2017.03.010](https://doi.org/10.1016/j.eswa.2017.03.010).

69. Talburt, John R. “1 - Principles of Entity Resolution.” *Entity Resolution and Information Quality*, edited by John R. Talburt, Morgan Kaufmann, 2011, pp. 1–37. *ScienceDirect*, doi:[10.1016/B978-0-12-381972-7.00001-4](https://doi.org/10.1016/B978-0-12-381972-7.00001-4).

70. OWL - Semantic Web Standards. <https://www.w3.org/OWL/>. Accessed 3 Mar. 2019.

71. SKOS Simple Knowledge Organization System Namespace Document - HTML Variant, 18 August 2009 Recommendation Edition. <https://www.w3.org/2009/08/skos-reference/skos.html>. Accessed 3 Mar. 2019.

72. *Linked Data - Design Issues*. <https://www.w3.org/DesignIssues/LinkedData.html>. Accessed 3 Mar. 2019.

the most comprehensive reference is the Virtual Information Authority File (VIAF), which “combines multiple name authority files into a single OCLC-hosted name authority service”⁷³.

The use of place name authorities is still however a topic discussion in the Cultural Heritage community. The most commonly used datasets are GeoNames and the Getty Thesaurus of Geographic Names (TGN), albeit with differing approaches. While it is generally agreed that GeoNames provides the greatest amount of coverage⁷⁴, TGN does provide added value as it records historical places as well. Following some random sampling for the the FotoIndex dataset, tests determined that many place names were missing from TGN, and since the place names referred to institutions from the 20th century, there was no need to reference historical places, and therefore GeoNames was selected as the principal authority file. Wikidata is another ever-expanding excellent source of data for these entities. Aside from providing structured data sources from Wikipedia, it contains an arsenal of identifiers from other more curated vocabularies, including ULAN, Geonames, and VIAF. Coverage is however an issue, as many of the vocabularies still require manual alignment with Wikidata, including ULAN which has roughly only 50% of the dataset aligned.⁷⁵

73. VIAF. <https://viaf.org/>. Accessed 3 Mar. 2019.

74. Acheson, Elise, et al. “A Quantitative Analysis of Global Gazetteers: Patterns of Coverage for Common Feature Types.” *Computers, Environment and Urban Systems*, vol. 64, July 2017, pp. 309–20. *ScienceDirect*, doi:[10.1016/j.compenvurbsys.2017.03.007](https://doi.org/10.1016/j.compenvurbsys.2017.03.007).

75. *Mix'n'match*. https://tools.wmflabs.org/mix-n-match/#/group/ig_art. Accessed 3 Mar. 2019.

	ULAN	AAT	VIAF	GeoNames	WikiData
Artist Name	✓				✓
Institution Type		✓			✓
Institution Name			✓		✓
Institution City				✓	✓

Figure 24: Terms matched to corresponding vocabularies

Due to the (currently) limited coverage of Wikidata, it was decided to align all entities to the best corresponding fit, in addition to WikiData. The table in figure 24 outlines the list of entities and their corresponding dataset.

Artist Names

The process and methodologies of reconciling one data set with another can vary greatly, depending on the quality of the source data and the level of coverage in the target dataset. In the case of the artist names, the process was fairly straightforward as all of the I Tatti datasets were originally based on ULAN. They did not however contain identifiers for ULAN, meaning that the only option was to perform a string match against the ULAN preferred or alternate name. This process can technically be achieved in many ways—less technically inclined individuals have even downloaded the entire dataset into Microsoft Excel and used a string matching

function. Alternatively, writing a Python script to query the Getty vocabulary endpoint is one another option, as outlined in a previous project at the Harvard Center.⁷⁶

ID	Artist	fourth	third	second	first result	FINAL
1	Tommaso	http://www.getty.edu/	http://www.getty.edu/	http://www.getty.edu/	http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
2	Abbate, Niccolò dell'			http://www.getty.edu/	http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
3	Abruzzo 14th-15th centuries				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
4	Acceptus				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
5	Agabiti, Pietro Paolo				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
6	Agnelli, Marino				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
7	Agnolo di Polo				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
8	Agostini, Giovanni Paolo de'				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
9	Agostino d'Antonio di Duccio				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
10	Agostino di Giovanni				http://vocab.getty.edu/ulan/5000459/	http://vocab.getty.edu/ulan/5000459/
11	Alamanno, Pietro				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
12	Albani, Francesco		http://www.getty.edu/	http://www.getty.edu/	http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
13	Alberegno, Jacopo				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
14	Alberti, Antonio			http://www.getty.edu/	http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
15	Alberti, Cherubino				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
16	Alberti, Durante				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
17	Albertinelli, Mariotto			http://www.getty.edu/	http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
18	Alberto Sotio				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
19	Alboresi, Giacomo				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
20	Alenis, Tommaso de				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
21	Aleotti, Antonio				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
22	Alesso d'Andrea				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
23	Alfani, Domenico di Paride				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
24	Alfani, Orazio				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
25	Aligardi, Alessandro				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
26	Allamagna, Justus d'				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
27	Allegretto di Nuzio				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
28	Allegri, Francesco			http://www.getty.edu/	http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
29	Allori, Alessandro			http://www.getty.edu/	http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
30	Allori, Cristofano				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
31	Altichiero			http://www.getty.edu/	http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
32	Alvaro di Piero				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
33	Amadeo, Giovanni Antonio				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
34	Amalteo, Pomponio				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
35	Ambrogio da Milano			http://www.getty.edu/	http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/
36	Ambrascini d'Acti				http://www.getty.edu/vow/ULANFull/	http://www.getty.edu/vow/ULAN/

Figure 25: reconciling artist names against ULAN

Another method involves running SPARQL queries against the Getty Vocab endpoint within OpenRefine⁷⁷, a common tool for performing data parsing, cleanup, and especially reconciliation. One must first fine-tune the query using the SPARQL endpoint, then convert the

76. Klic, Lukas, et al. *The Code4Lib Journal – The Drawings of the Florentine Painters: From Print Catalog to Linked Open Data*. no. 38, Oct. 2017, <https://journal.code4lib.org/articles/12902>.

77. *Getty Vocabularies LOD: Sample Queries*. http://vocab.getty.edu/queries#OpenRefine_Reconciliation_Service. Accessed 3 Mar. 2019.

query to a format that OpenRefine will be able to utilize and return results in JSON format. Once those results are returned they must be parsed again to extract the identifier and corresponding label that one is searching. The Getty vocabularies primarily use a complex combination of SKOS and a custom ontology as their data model, making the web of relationships difficult to navigate for users who are not familiar with their structure. In order to provide a more streamlined access to some of the key data points (such as labels for artist names), they do provide a full text index that allows one to query it in a rather simple way to arrive at very useful results.

```
select ?x ?label {
  ?x luc:term "michelangelo buonarotti";
  gvp:prefLabelGVP/xl:literalForm ?label. }
```

This sample query will return a list of matching terms with good relevancy ranking, leveraging the Lucene index (using the keyword “luc:term” as predicate) in the graph database backend (GraphDB⁷⁸ in this case) to search across preferred or alternate names of artists. This Lucene index provides much better relevancy ranking than a traditional SPARQL query, which generally has little tolerance for possible variances in spelling or formatting, and requires a regular expression for more complex string searching functions. As with most reconciliation activities, the objective is to obtain a solid balance of precision and recall for the results. Given the lack of additional contextual data about artists in the FotoIndex dataset (such as their nationality, date of birth, etc), and the occasional overlap in the names of different artists, it was

78. “Ontotext GraphDBTM - a Semantic Graph Database Free Download.” *Ontotext*, <https://www.ontotext.com/products/graphdb/>. Accessed 4 Mar. 2019.

not possible to properly disambiguate the returned entities from the Getty vocabularies in a fully automated manner.

The results from the entity reconciliation process were subsequently loaded into a Google Sheet for sharing and analysis by subject experts. Out of 1768 artist records contained in the FotoIndex dataset, 1453 (82%) had corresponding matches in the Getty ULAN. 235 (13%) had two or more matches, and only two records had four or more matches. It was therefore decided to keep the top four results for subsequent analysis.

Following a process of manual checking a random selection of the records that returned only a single result, it was determined that the quality of linking was reliable enough and it would not be necessary to manually check all records. Manual linking would need to be performed on the 235 records that returned more than one result, and the 233 records that yielded no results, corresponding to roughly 24% of the total records. This process was performed directly in the Google Sheet by subject experts, by constructing links that passed the name of the artist to search directly into the web search interface of ULAN.

Following a manual alignment process performed by the subject expert, it was found that of the 18% of the total records did not return any result, most of them were records for “unknown” Italian artists that referenced a particular region in Italy rather than a specific person. These records were specific to I Tatti collections so it was expected that they would not be present in the ULAN dataset. These records, which usually contained a geographic region and time period (such as “Abruzzo 14th-15th centuries”), were instead given a GeoNames URI for the location and a date range for the time period. Later, as these would be published as more

structured, linked data, it would allow for the retrieval of these artworks by region and time period.

Of the 235 records that returned more than one result during the reconciliation process, it was found that over half (120) had duplicate entries in ULAN. This was the result of many records being loaded from disparate datasets at the Getty, where a full de-duplication process had not been performed. These records have been since merged and corrected, so if the process were performed again in 2019, and if one were to exclude the anonymous Italian artists, the success rate for the reconciliation process would be at 93%.

Following the linking to ULAN, SPARQL queries were run against Wikidata using the ULAN identifiers in order to build sameAs links between these records. This query resulted in 1217 results, which represents a 79% success rate, far higher than the 51% general coverage between the two datasets that is reported by the Mix'n'match tool from the Wikimedia foundation.⁷⁹ The discrepancy here is most likely due to the fact that the artists in the FotoIndex collection are historical actors that have received substantial attention from publications, whereas many entries in ULAN are for contemporary artists who have had little written about them and may not have WikiData entries.

Finally, in order to create links to internal databases for additional alignment and metadata enrichment, a string match was run against the “Mappatura” internal database that contains an inventory of the physical layout of collection, links to existing digitized records, and collection-level records by artist stored in the bibliographic database HOLLIS.

79. *Mix'n'match*. https://tools.wmflabs.org/mix-n-match/#/group/ig_art. Accessed 3 Mar. 2019.

All of these links greatly enrich the overall quality of the dataset, allowing for contextual information to be retrieved for artists, such as their gender, date of birth, nationality, and region in which they were active. Additionally, and possibly the most useful for art historians is data about relationships to other artists, such as teacher-pupil, father-son, follower, school of, etc. An overview of the results of the alignment work can be seen in Figure 26.

Total FotoIndex Artist Records	1768	100%
ULAN Matches	1535	86%
ULAN Matches (more than one)	235	13%
ULAN duplicate entries	120	7%
Unknown Italian	182	10%
Wikidata Links	1217	68%
Links to “Mappatura” internal database	1438	81%
Links to HOLLIS records	1034	58%

Figure 26: Matching statistics for artist names

Provenance Records

Provenance data, which included the institution name and geographic location derived from transcriptions of handwritten annotations on the backs of the photographs, proved to be far more complex a task for entity alignment. Each photograph contained the name of the city where the artwork was held, along with a corresponding institution. Although the location of the

artwork may have changed since scholars made this annotation, the information was deemed as being important for provenance research. The transcriptions of these annotations in list form were included in the original index records created by the Getty in the 1980's, and were generally well-formatted, consistent, and accurate, with delimiters that allowed for the splitting of data into corresponding tables. As a result of the initial Matchmaker process linking the images to the provenance data, nearly 12,000 records needed to be reconciled to external vocabularies.

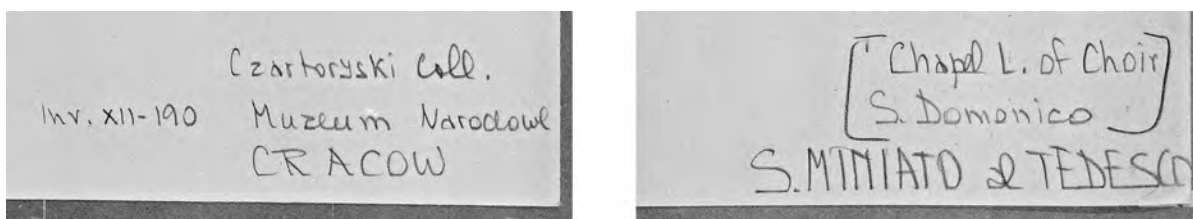


Figure 27: provenance information on verso of photographs

The challenge was with the “organic” nature of the source data that did not lend itself well to entity reconciliation. Large institutions such as the Uffizi Gallery or Metropolitan Museum were fairly straightforward entities to work with, as they have authority records in most datasets, including Wikidata and VIAF. This provenance data included records for museums (and collections within them), churches, libraries, private collections, collectors, monuments, public works, architecture, among many others. Given the many duplicates and vague descriptions in various languages on the original photograph, a first attempt at reconciling these records using the Google Knowledge Graph API was performed, with the hopes that it would be possible to leverage relevancy ranking, autocorrect, and other functionality built by Google over the years.

At the time of testing, the API provided different results from those that are typically returned on the right side of the window of a simple Google search (as seen in figure 28).

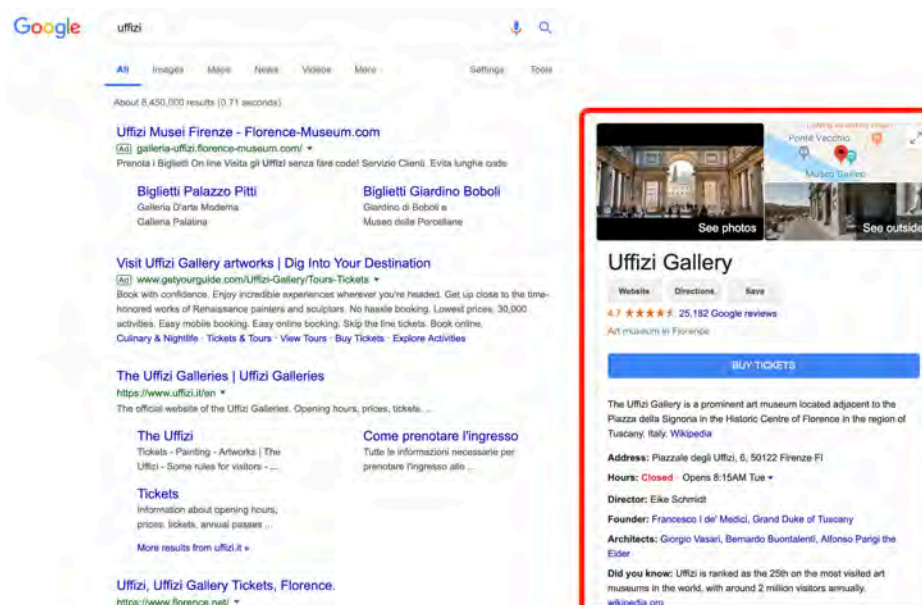


Figure 28: Google Knowledge Graph structured data returned from ambiguous search.

From a series of tests performed with some random samples of data, along with tests of collection names that were known to be more obscure, traditional google searches proved to be very effective. A Google search for “Uffizi” would display data from the Google Knowledge Graph in the side panel, achieving accurate results on nearly all tests performed. Many of the records, particularly those that were more obscure, did not show any results from the Google Knowledge Graph however, especially if there was no corresponding record in Wikipedia. Issues arose however when running these same queries programmatically using the API, as there was no autocorrect for variances in the spelling of names, and it generally was not able to handle the level of fuzziness in the source records. The intent was that with the Google Knowledge Graph API results, Wikipedia links could be extracted from the results, which could then be connected

to WikiData, Google Maps, GeoNames, and VIAF, all programmatically without user intervention. Links to the Google Knowledge Graph entity itself were not of interest, as the dataset is closed and is generally not used for integration methods across repositories. The hope was instead to leverage the disambiguation and relevancy ranking functionality that Google offered. At the time it was found that direct queries to the target services proved to be more effective, as those services returned more than one result for a given string query with additional contextual data (such as the type of entity in the case of VIAF and WikiData). Additionally, with services such as VIAF that contain many duplicates, it was important to have more than one result. Additionally, VIAF categorizes entities based on type, such as “Corporate”, “Geographic”, and “Personal”. Problems with this structuring emerged particularly with private collections, where a notable individual had a collection of artworks, and records were found for the individual but not the collection itself. For these types of records, sameAs links would not be ontologically accurate with a single data model that makes a statement about the work being held in a collection, since the type of the entity was a person rather than a collection.

Following this assessment of existing records, it was determined that it would be best to build a reconciliation module for the MatchMaker application that would allow subject experts with domain knowledge to review the results returned from the various external datasets and link them to the internal institutional record. A first pass with the reconciliation service in OpenRefine allowed for the extraction of links in GeoNames, WikiData, and VIAF. Publicly available modules and services built for OpenRefine, such as the “Conciliator” app available on GitHub, allow one to quickly plug into to the VIAF service.⁸⁰ Reports similar to those published

80. Chiu, Jeff. *OpenRefine Reconciliation Services for VIAF, ORCID, and Open Library + Framework for Creating More.: Codeforkjeff/ Conciliator*. 2016. 2019. *GitHub*, <https://github.com/codeforkjeff/conciliator>.

by the Smithsonian Libraries⁸¹ on reconciling their collections data proved to be very useful as a guide to navigating the process, as the process requires some fine tuning. Small modifications to restrict or widen search parameters can return substantially different results-- for example the ability to add contextual data for the lookup, such as organization type or location can produce a lower recall but higher precision. Following a number of tests, it was determined that adding contextual information to the search parameters produced a set of results that was too restrictive, in particular for entities that were more obscure.

The process began by loading up the full dataset for the institutional records, and separate columns were created for matching results from VIAF using the Conciliator plugin, GeoNames using the GeoNames reconciliation service⁸², and WikiData was queried though SPARQL queries that were constructed in similar ways to ULAN.

81. Ota, Allyson. *Reconciling Smithsonian Library Data with VIAF*. Smithsonian Libraries, 8 Sept. 2016.

82. Harlow, Christina. *GeoNames Reconciliation Service for OpenRefine/LODRefine/Google Refine: Cmharlow/Geonames-Reconcile*. 2015. 2019. *GitHub*, <https://github.com/cmharlow/geonames-reconcile>.

11896 rows

Show as: rows records Show: 5 10 25 50 rows

All	ID	Location	Institute	recon3	recon2	recon1	reconcile
	1	2613 (Portugal) S. Bras de Alportel Create new topic	Don Luis Bramão Create new topic				Don Luis Bramão
	2	2614 Cave Cave 18 21019, -78.0407 (100) Cave 44 30935, 170 95009 (100) Cave 41 81682, 12 94055 (100) Create new topic	Chiesa della Cona (Umbrian XVI c.) Create new topic				Chiesa della Cona (Umbrian XVI c.)
	3	2615 A cave (Umbria) Create new topic					
	4	2616 Aalen (Württemberg) Create new topic	Schloss Fachsenfeld Sammlung Schloss Fachsenfeld (0.055) Stiftung Schloss Fachsenfeld (0.593) Schloss Fachsenfeld (0.05) Create new topic	Schloss Fachsenfeld,240543518	Stiftung Schloss Fachsenfeld,262017152	Sammlung Schloss Fachsenfeld,150076664	Schloss Fachsenfeld
	5	2617 Abbazia S. Salvatore Create new topic	Abbazia Cistercense Monasterio de Las Huelgas de Burgos (Spain) (0.162) Abbaye des Dunes (Brijuni) (0.22) Zisterzienserstift Stams (0.167) Create new topic	Zisterzienserstift Stams,130809091	Abbaye des Dunes (Brijuni),140949010	Monasterio de Las Huelgas de Burgos (Spain),140219949	Abbazia Cistercense
	6	2618 Abondance (Savoia) Create new topic					
	7	2619 Acrenza (Basilicata) Create new topic	Duomo Catholic Church, Archdiocese of Salzburg (Austria) (0.061) Biblioteca capitolare di Verona (0.065) duomo di Milano (0.267) Create new topic	duomo di Milano,151230415	Biblioteca capitolare di Verona,157684877	Catholic Church, Archdiocese of Salzburg (Austria),305788823	Duomo
	8	2620 Ackland, N.C. Create new topic	Art Museum Museum Yitra el (0.067) Metropolitan museum of art New York, N.Y. (0.165) Kunsthistorisches Museum Wien (0.267) Create new topic	Kunsthistorisches Museum Wien,126584224	Metropolitan museum of art New York, N.Y.,126238294	Muzeon Yitra el,141391944	Art Museum
	9	2621 Acquacanna Acquacanna 43.02859, 13.176 (100) Create new topic	S. Margherita in Valle Canto Create new topic				S. Margherita in Valle Canto
	10	2622 Acquapendente Acquapendente 42.74259, 11.86827 (100) Acquapendente 42.74383, 11.86418 (100) Create new topic	Duomo Catholic Church, Archdiocese of Salzburg (Austria) (0.06) Biblioteca capitolare di Verona (0.065) duomo di Milano (0.267) Create new topic	duomo di Milano,151230415	Biblioteca capitolare di Verona,157684877	Catholic Church, Archdiocese of Salzburg (Austria),305788823	Duomo

Figure 29: provenance data being reconciled in OpenRefine

As seen in figure 29, the results provided a solid foundation for moving forward with entity reconciliation, as most of the entries contained multiple responses from the various providers. Aside from matching data to that of external data sets, it became more and more clear that a substantial amount of cleaning would need to be performed as well.

ID	Location	geo3	geo2	geo1	Institute	recon3
2613	(Portugal) S. Bras de Alportel				Don Luis Bramão	
2614	Cave	Cave 41.81682	Cave -44.30835	1 Cave 18.2101	Chiesa della Cona (Umbrian XVI c.)	
2615	A cave (Umbria)					
2616	Aalen (Württemberg)				Schloss Fachsenf	Schloss Fachsenfeld,,2405435
2617	Abbadia S. Salvatore				Abbazia Cistercer	Zisterzienserstift Stams;13080
2618	Abondance (Savoia)					
2619	Acerenza (Basilicata)				Duomo	duomo di Milano;151230415
2620	Ackland, N.C.				Art Museum	Kunsthistorisches Museum Wi
2621	Acquacarina			Acquacarina	S. Margherita in Valle Canto	
2622	Acquapendente		Acquapendente 4	Acquapendent	Duomo	duomo di Milano;151230415
2623	Acquapendente		Acquapendente 4	Acquapendent	S. Pietro	Ecclesia Catholica. Pontificalia
2624	Acquapendente (Viterbo)				S. Francesco	Tertius ordo regularis Sancti Fr
2625	Acqui	Forte Acqui 44	Acqui Terme 44.6	Acqui Terme	Duomo	duomo di Milano;151230415
2626	Adelaide				National Gallery	National gallery of Australia;12
2627	Aerdenhout			Aerdenhout 5	Dufour Collection	
2628	Agira (Enna)				S. Maria Latina	
2629	Agira (Enna)				S. Salvatore	Societas Divini Salvatoris.;149E
2630	Agnano (S. Giuliano Terme)				Parish Church	St. Joseph's Parish Zell, Mo;14
2631	Agnano (S. Giuliano Terme)				S. Jacopo	San Iacopo di Ripoli (convento
2632	Agolla (Pioraco)					
2633	Agordo	Canale d'Agordo	Agordo 46.27992	Agordo 46.2E	Chiesa Arcipretale	
2634	Agrigento	Agrigento Cathed	Agrigento 40.403C	Agrigento 37.31065	13.57661; http://sws.geonames.org/252	
2635	Agrigento	Agrigento Cathed	Agrigento 40.403C	Agrigento 37	Cathedral	Westminster Abbey.;13246084
2636	Agrigento	Agrigento Cathed	Agrigento 40.403C	Agrigento 37	Museo	Mathaf al-misri;123996097
2637	Agrigento	Agrigento Cathed	Agrigento 40.403C	Agrigento 37	S. Maria dei Greci	
2638	Agrigento	Agrigento Cathed	Agrigento 40.403C	Agrigento 37	S. Rosalia	Centro de Estudios Rosalia de I
2639	Agrigento	Agrigento Cathed	Agrigento 40.403C	Agrigento 37	S. Spirito	Ospedale Santo Spirito in Sassi
2640	Aigueperse	Aigueperse 45	Aigueperse 46.27E	Aigueperse 4	Notre Dame	Trappists;127431364
2641	Aix en Provence	Kyriad - Aix En Pr	Aix-en-Provence 4	Aix-en-Provenc	Musée	Gosudarstvennaja Tret'jakovsk
2642	Aix-en-Provence	Kyriad - Aix En Pr	Aix-en-Provence 4	Aix-en-Provenc	Estang de Parade	Family
2643	Aix-en-Provence	Kyriad - Aix En Pr	Aix-en-Provence 4	Aix-en-Provenc	Musée	Gosudarstvennaja Tret'jakovsk
2644	Aix-en-Provence	Kyriad - Aix En Pr	Aix-en-Provence 4	Aix-en-Provenc	Musée (now Paris: Musée Nationalux)	
2645	Ajaccio	Ajaccio Canyon	Ajaccio 41.92556	Ajaccio 41.91886	8.73812; http://sws.geonames.org/3038334	
2646	Ajaccio	Ajaccio Canyon	Ajaccio 41.92556	Ajaccio 41.91	Fesch Collection	
2647	Ajaccio	Ajaccio Canyon	Ajaccio 41.92556	Ajaccio 41.91	Musée	Gosudarstvennaja Tret'jakovsk
2648	Ajaccio	Ajaccio Canyon	Ajaccio 41.92556	Ajaccio 41.91	Musée d'Ajaccio	

Figure 30: provenance data issues

As can be seen in figure 30, issues with duplicate records are highlighted in blue and records lacking institution names or other ambiguities in green. The complexity of these issues made it evident that it would be necessary to build a custom module for the MatchMaker application that would allow cataloging staff to go through the records in a single pass, and systematically perform all of the cleanup and linking necessary to make the data more useful and interoperable. Unlike the original matching activities that did not require any research on the entity, the process of looking up some institutions or private collectors was substantially more time consuming. A focus was again kept on attempting to maintain a balance of capturing the

most amount of useful data, while creating a process that would not take too much time to complete. While reducing the number of clicks to link a record for an institution was kept as a key objective, the records that had to be crosschecked were on external websites, so that process required the application to open browser windows that would display these records.

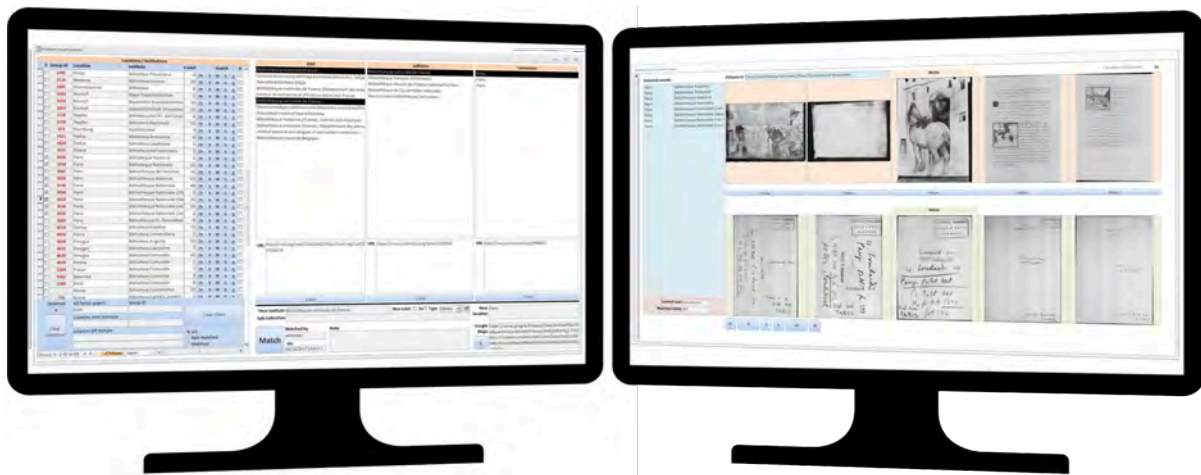


Figure 31: provenance data matching interface

Leveraging much of the same functionality from the original Matchmaker application and the lessons learned from those workflows, a similar user interface was built using two screens.

After some deliberation and consultation with staff from the photographic archive, it was

determined that the following tasks should be performed for each record:

- Merge duplicate institutional records
- Align the institution name with VIAF and WikiData
- Align the city name with GeoNames
- Create a standardized name for the institution in English, using the Library of Congress authority record, when possible
- Identify the institution type (Museum, Private Collection, Church, Library, etc.)
- Locate the institution on Google maps and provide a link to the Google places record.

The screenshot displays the 'location-reconciliation' interface. On the left, a table lists various institutions with columns for Group ID, Location, Institute, Count, Search, and R. A vertical green highlight covers several rows, including those for Paris (Group IDs 3658, 3659, 3660, 3661, 3662, 3663, 3664, 3665, 3666, 3667, 3668, 3669, 3670, 3671, 3672, 3673, 3674, 3675, 3676, 3677, 3678, 3679, 3680, 3681, 3682, 3683, 3684, 3685, 3686, 3687, 3688, 3689, 3690, 3691, 3692, 3693, 3694, 3695, 3696, 3697, 3698, 3699, 3700, 3701, 3702, 3703, 3704, 3705, 3706, 3707, 3708, 3709, 3710, 3711, 3712, 3713, 3714, 3715, 3716, 3717, 3718, 3719, 3720, 3721, 3722, 3723, 3724, 3725, 3726, 3727, 3728, 3729, 3730, 3731, 3732, 3733, 3734, 3735, 3736, 3737, 3738, 3739, 3740, 3741, 3742, 3743, 3744, 3745, 3746, 3747, 3748, 3749, 3750, 3751, 3752, 3753, 3754, 3755, 3756, 3757, 3758, 3759, 3760, 3761, 3762, 3763, 3764, 3765, 3766, 3767, 3768, 3769, 3770, 3771, 3772, 3773, 3774, 3775, 3776, 3777, 3778, 3779, 3780, 3781, 3782, 3783, 3784, 3785, 3786, 3787, 3788, 3789, 3790, 3791, 3792, 3793, 3794, 3795, 3796, 3797, 3798, 3799, 3800, 3801, 3802, 3803, 3804, 3805, 3806, 3807, 3808, 3809, 3810, 3811, 3812, 3813, 3814, 3815, 3816, 3817, 3818, 3819, 3820, 3821, 3822, 3823, 3824, 3825, 3826, 3827, 3828, 3829, 3830, 3831, 3832, 3833, 3834, 3835, 3836, 3837, 3838, 3839, 3840, 3841, 3842, 3843, 3844, 3845, 3846, 3847, 3848, 3849, 3850, 3851, 3852, 3853, 3854, 3855, 3856, 3857, 3858, 3859, 3860, 3861, 3862, 3863, 3864, 3865, 3866, 3867, 3868, 3869, 3870, 3871, 3872, 3873, 3874, 3875, 3876, 3877, 3878, 3879, 3880, 3881, 3882, 3883, 3884, 3885, 3886, 3887, 3888, 3889, 3890, 3891, 3892, 3893, 3894, 3895, 3896, 3897, 3898, 3899, 3900, 3901, 3902, 3903, 3904, 3905, 3906, 3907, 3908, 3909, 3910, 3911, 3912, 3913, 3914, 3915, 3916, 3917, 3918, 3919, 3920, 3921, 3922, 3923, 3924, 3925, 3926, 3927, 3928, 3929, 3930, 3931, 3932, 3933, 3934, 3935, 3936, 3937, 3938, 3939, 3940, 3941, 3942, 3943, 3944, 3945, 3946, 3947, 3948, 3949, 3950, 3951, 3952, 3953, 3954, 3955, 3956, 3957, 3958, 3959, 3960, 3961, 3962, 3963, 3964, 3965, 3966, 3967, 3968, 3969, 3970, 3971, 3972, 3973, 3974, 3975, 3976, 3977, 3978, 3979, 3980, 3981, 3982, 3983, 3984, 3985, 3986, 3987, 3988, 3989, 3990, 3991, 3992, 3993, 3994, 3995, 3996, 3997, 3998, 3999, 4000).

The right side of the interface shows a detailed view of the 'Bibliothèque nationale de France' (BnF) entry. It includes a map of Paris, a list of associated URLs, and a 'Match' section. The 'Match' section shows a match between the 'New Institute' (Bibliothèque nationale de France) and the 'Sub Collection' (Bibliothèque nationale de France). The 'Match' section also includes a 'Matched by' field (addressen) and a 'Note' field (30/10/2017 15:51:37). A red box highlights the 'Match' section, and another red box highlights the 'Google Maps' link in the bottom right corner.

Figure 32: detail of provenance matching interface

While the list of tasks became rather extensive, the opportunity it would provide scholars doing provenance research in the future was significant. It was anticipated that the ability to group duplicate records together (which were not de-duped during the original matching process), as seen highlighted in green in figure 32, would substantially reduce the number of records that would need to be researched.

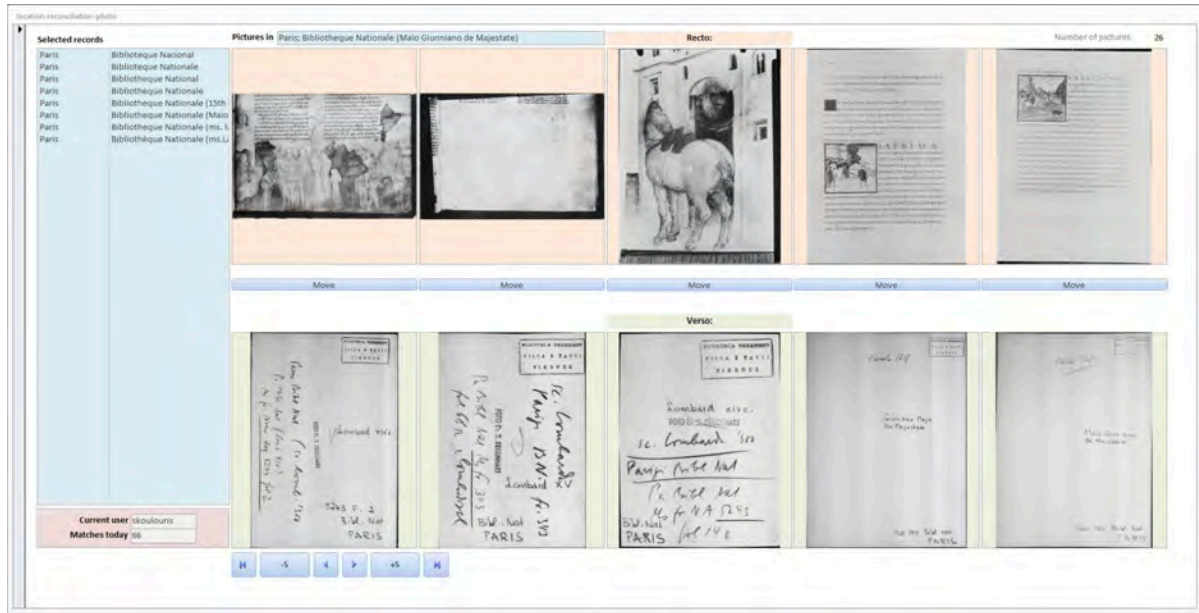


Figure 33: detail of provenance matching interface (right screen)

As illustrated in figure 33, the second screen was used to display the photographs that are associated with the selected institution record, allowing the user to browse through the images to visually confirm on the verso of the photographs that the collection was in fact the same in instances where multiple records were selected.

The screenshot displays a 'location-reconciliation' interface. On the left, a table lists various institutions with columns for 'Group ID', 'Location', 'Institute', 'Count', and 'Search'. The table includes entries for locations like Milan, Montecassino, Munich, and various libraries in France and Italy. Below the table are search filters and a 'Clear selected' button. On the right, three panels show data from external sources: 'VIAF', 'Wikidata', and 'Geonames'. The 'VIAF' panel lists several national libraries. The 'Wikidata' panel shows a map of France. The 'Geonames' panel shows a map of Paris. At the bottom, a 'Match' section displays a Google Maps link and a '6' in a red circle, indicating a match count.

Figure 34: detail of provenance matching interface (right screen)

Various buttons, as seen in figure 34, provided quick lookup functionality in a web browser to our external sources by building the URL from existing data. The dual-screen setup, along with an optimized user interface, allowed for the enrichment of these 12 thousand rather chaotic records in a relatively efficient manner with human-level precision. Catalogers did however often find themselves stuck for extended periods of time researching private collectors or poorly documented churches and monuments in Italy, which made the reconciliation and data cleanup process a much more arduous task than originally anticipated.

A quantitative summary of the enrichment efforts can be seen in Figure 35. Of the original 11,896 records, 45% of these were de-duplicated by this process. At the outset, the possibility of automating the de-duplication process was investigated but it was found that too many of the records provided vague information that required the interpretation of domain experts. For example, a record listed as “Aix-en-Provence” is the name of a city, but domain

experts could infer that since there was only one museum in the city, the reference was to that institutional record. Additionally, the fuzziness of multiple records could provide additional context to the user that could assist in the selection process. As argued by Trevor Muñoz in his article “Against Cleaning”⁸³ this context is important to scholars who are viewing these records and should be preserved in the final records. As demonstrated by the author in the article “Florentine Renaissance Drawings: a Linked Catalog for the Semantic Web”⁸⁴, this context can be preserved by maintaining the source content and leveraging Linked Data technology to facilitate the data integration layer.

Institution Records Uncleaned	11,896	
De-duplicated institution records	5351	100%
VIAF Corporate	2517	47%
VIAF Geographic	828	15%
VIAF Personal	37	1%
Geonames	5319	99%
Wikidata Links	2510	47%
Google Maps	3989	75%

Figure 35: Matching statistics for institutional records

83. Rawson, Katie, and Trevor Muñoz. *Against Cleaning*. July 2016. *curatingmenus.org*, <http://www.curatingmenus.org/articles/against-cleaning/>.

84. Klic, Lukas, et al. “Florentine Renaissance Drawings: A Linked Catalog for the Semantic Web.” *Art Documentation: Journal of the Art Libraries Society of North America*, vol. 37, no. 1, Mar. 2018, pp. 33–43. *www-journals-uchicago-edu.ezp-prod1.hul.harvard.edu (Atypon)*, doi:[10.1086/697276](https://doi.org/10.1086/697276).

As with the image matching process, the application captured the amount of time spent on each record. Although it was anticipated that a large portion of the records could be matched quickly and easily, given the number of small churches and private collections, there were many uncertainties with regards to the amount of time it would take to research these smaller collections. Initially it was calculated that the task could be completed within three months with three staff dedicating a few hours a day. After the first week of data cleanup, the statistics showed that the process was moving along far slower than anticipated. It was decided to dedicate five staff to the process, and create benchmarks with a goal of fifty records per day per staff member.

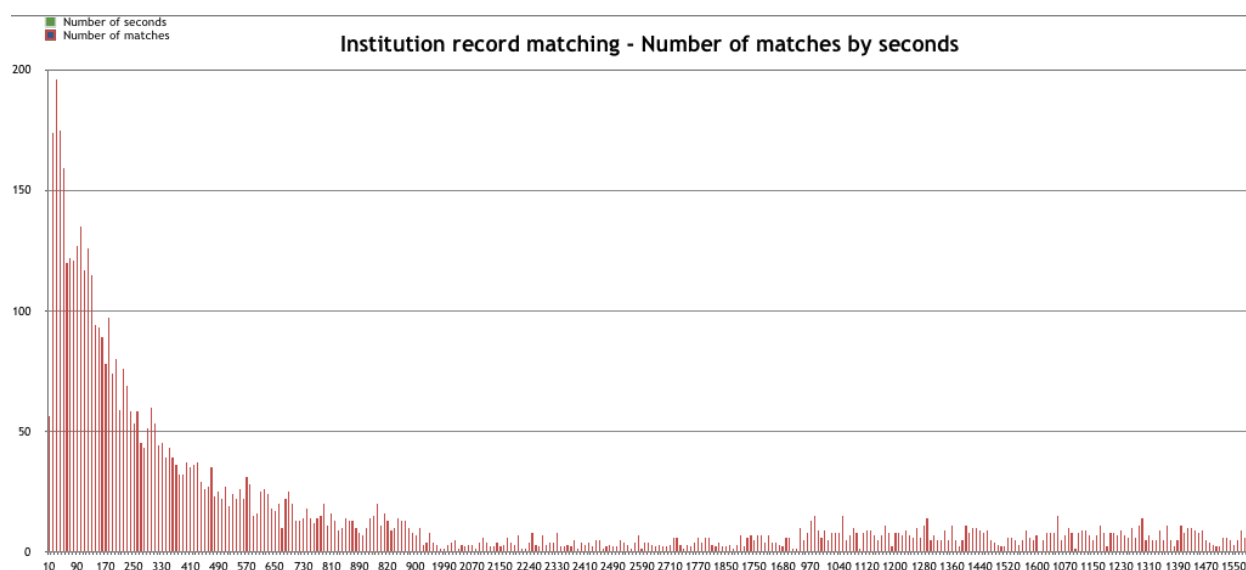


Figure 36: statistics on provenance matching tasks

At the end of the process it was seen that for a single record, it took anywhere between 20 seconds to 45 minutes to do the research, a range that is far higher than anyone could have anticipated. Although 85% of the matching tasks were completed in under 15 minutes (900 seconds), this represents only 35% of the total time spent to complete all records. As seen from

figure 36, there was some concentration in the range between thirty seconds to two minutes, but the greatest portion of these were spread out all the way up to forty-five minutes. The resulting work ended up in over five thousand unique records, with an average of ten minutes per record.

Enrichment Outcomes

While the original image matching process took a total of 223 hours of matching work, the provenance cleanup and alignment took 874 hours. These figures do not include all of the preparation work, which included initial entity reconciliation, preparatory data cleanup, and software development for the matchmaker application. Although the provenance data cleanup was disproportionately higher, the quality of the data is very high and will provide a unique viewpoint and can serve as an extraordinary resource for research on the distribution of Early Modern artworks from the Italian peninsula during the middle of the twentieth century.

The screenshot shows a 'Collection-level records' view. On the left, there is a list of attributes: Artist name, Number of images and location on the reel, and List of institutions (provenance). On the right, there is an XML snippet representing a record.

```

<record>
  <recno>3230</recno>
  <Project_neg_182 (354-388)</Project_neg_>
  <No_of_images>35</No_of_images>
  <COUNT>35</COUNT>
  <I_Tatti_ref_XXV.20s</I_Tatti_ref_>
  <Artists_names>Dossi, Dossa</Artists_names>
  <Object_locs>Rome: Castel S. Angelo; Campidoglio; Galleria Nazionale Barberini; Galleria Borghese
  <Object_locs-Vatican City: Pinacoteca Vaticana</Object_locs>
</record>

```

Figure 37: view of a collection-level record

As can be seen in Figure 37, the FotoIndex dataset began with group records that would have been of little use by scholars. Through the cleaning, enrichment, and reconciliation process,

this dataset was transformed to highly useful collection of normalized and reconciled data with URI's linking to other resources that provide additional data on each entity (see Figure 38).

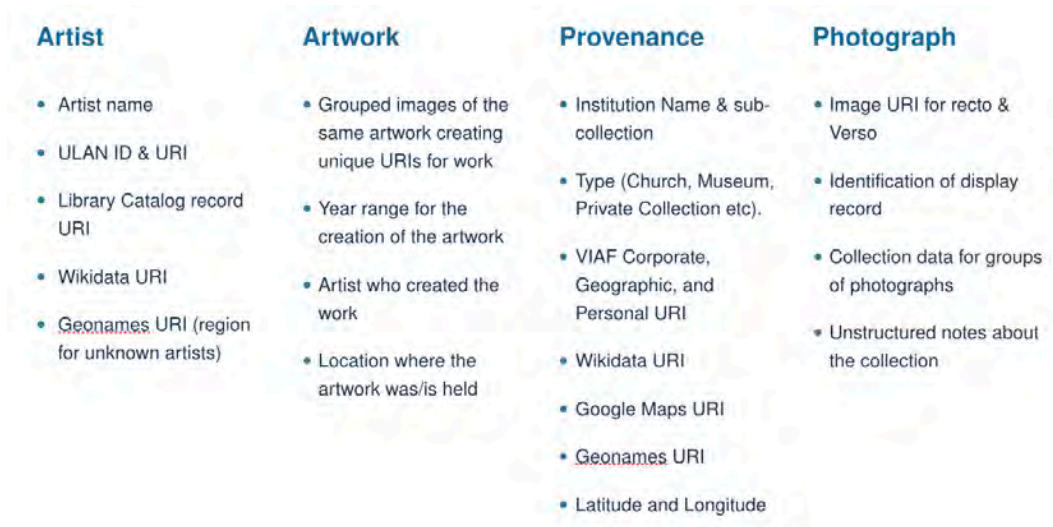


Figure 38: overview of data collection and enrichment results

The amount of data that was generated for the images should be sufficient for scholars to be able to find images for specific artworks, gain a larger overview about the history of collecting, study the history of attributions, and be able to access the annotations and research data through images of the backs of the photographs. Access to the image itself, will also open new points of access when leveraging and building new tools that employ computer vision and machine learning methodologies. The next chapter will outline how to use this rich foundation of data to transform it to usable, 5-star linked data for publishing on the Semantic Web.

IV - Synergizing and Publishing Data

Options for publishing collections of cultural heritage material are quite numerous, as the past two decades have seen much development in this front. Emerging open-source solutions have proven to be the most sustainable for cultural heritage institutions, as one is not bound to a particular vendor or data format, and can extract the data out of the system at any point. This last point has been particularly painful for most large museums around the world who chose TMS⁸⁵ as their collections management suite, a commercial product from GallerySystems. The business models of these companies have proven to be very frustrating for museums, as they do not provide any mechanism for direct and open access to the data, even for the institution itself, without paying fees which can be prohibitive. These problems have driven many institutions towards open-source solutions, which have in turn resulted in many excellent projects. One such example is the Omeka platform from the Roy Rosenzweig Center for History and New Media at George Mason University, which provides solutions to quickly take collections data and publish them in the digital form.⁸⁶ With the recent release of Omeka S, the option to publish using any ontology has become possible, as well as making all of the data available through an API serialized as JSON-LD. Another open-source product, Project Blacklight⁸⁷, is a Ruby on Rails-based application that has seen substantial development over the years, including a number of extensions that allow collection curators to log in and manage data, as well as to add support

85. "TMS Collections Collection Management Software, Museum Collections." *Gallery Systems*, <https://www.gallerysystems.com/products-and-services/tms-suite/tms/>. Accessed 8 Mar. 2019.

86. *Omeka*. <https://omeka.org/>. Accessed 8 Mar. 2019.

87. *Blacklight*. <http://projectblacklight.org/>. Accessed 8 Mar. 2019.

for geospatial metadata and collections of historical maps. The core Blacklight version runs off of a SOLR index⁸⁸ so it is not meant to manage the data as a traditional collections management systems does, but rather provide the tools to quickly index and retrieve data. It does however, provide an extremely intuitive and responsive user interface that can read from other data stores, allowing multiple collections of metadata to be indexed and connected. Blacklight was used in an earlier digital edition project, *The Drawings of the Florentine Painters*⁸⁹ at the Harvard Center, providing a front end to RDF data. By running a SPARQL query against the data endpoint, a rake task parsed the results and stored them in a SOLR index that was preconfigured on that data set.⁹⁰

While these solutions are not RDF-native applications, they do provide simpler solutions that can be implemented quickly and with less overhead if the objective is to simply publish data and make it accessible. Despite the relative maturity of Linked Open Data and Semantic Web technologies, there are few open-source software solutions that present this technology in ways that allow individuals who are not data science professionals to effectively make use of it. Google, with its acquisition of Freebase and construction of the Google Knowledge Graph have managed to make the technology available at a large scale in their search engine.

88. Apache Solr -. <https://lucene.apache.org/solr/>. Accessed 8 Mar. 2019.

89. *The Drawings of the Florentine Painters*. <http://florentinedrawings.itatti.harvard.edu/>. Accessed 8 Mar. 2019.

90. *Solr 5.5 Config Files for Florentine Drawings Project: Villaitatti/Florentine-Drawings-Solr-Config*. 2016. Villa I Tatti | The Harvard University Center for Italian Renaissance Studies, 2018. *GitHub*, <https://github.com/villaitatti/florentine-drawings-solr-config>.

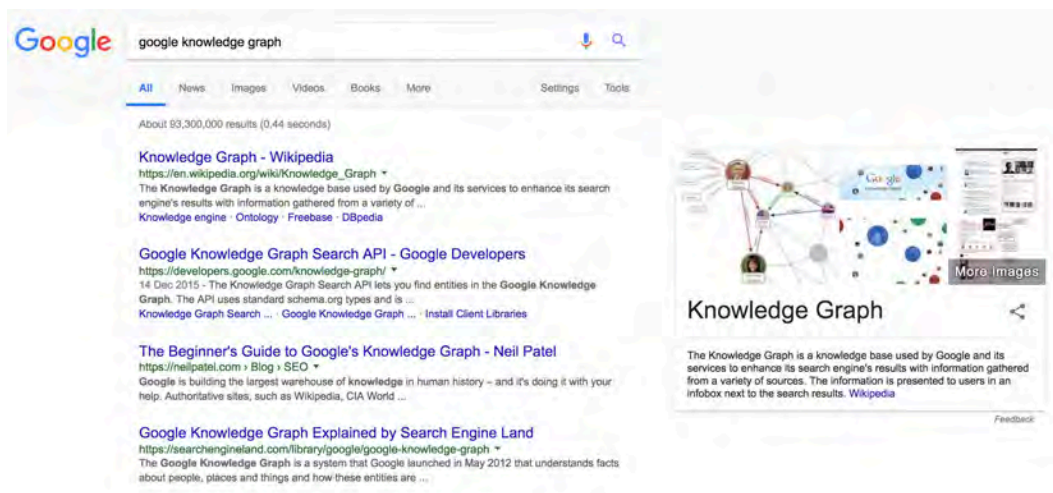


Figure 28: Google Knowledge Graph panel

As briefly outlined in the previous chapter, results are visible in the panel on the right when viewing results on the search results page, information which is constructed entirely from semantically enriched statements scraped from the web, structured by SKOS. While many organizations are publishing semantically enriched Linked Data, Google is one of the few who have managed to consume and republish that data in a meaningful way.

The meaningful consumption of semantically enriched Linked Data is in part, a question of scale. Many technology giants are now amassing large quantities of data in order to provide solutions to voice assistants (Google, Amazon, Apple) and provide more meaningful results to user questions, but these are operating at a scale that is beyond the capacity of the cultural heritage sector, and have an entirely different use-case. In order to be able to obtain meaningful results from the cultural heritage sector, a great deal of coordination among organizations is still necessary. Despite an ever-increasing number of conferences and symposia on Digital Humanities initiatives, the field is still lacking in widespread agreement on many standards and modes of implementation. Until the community is able to find more agreement on standards and

software implementations, the ability to ask complex questions that scale in the cultural heritage domain will be limited.

Linked Data Transitions

The vast majority of institutions that chose to publish Linked Data, still prefer to keep legacy systems, building Extraction Transformation and Load (ETL) processes to convert data held in traditional relational databases to RDF, the format of Linked Data. This transformed data is then stored in a graph database or tripplestore, and made accessible to developers through a SPARQL endpoint, API's, data dumps, or a combination of the three. Any data alignment with other datasets needs to be encoded and materialized directly in the source dataset (such as URIs for entities in other vocabularies), or in a separate (alignment) dataset that also gets integrated during the ETL process, by reconciling internal entities with external ones through owl:sameAs statements. This results in a cumbersome data management process, as institutions must manage both legacy systems in relational databases and RDF-based systems. Additionally, access to RDF data is limited to access through API's, where any potential enrichment from the user would require this data to go through a similar ETL process to be contributed back to the source dataset.

If the cultural heritage community has generally agreed that Linked Open Data, RDF, and the Semantic Web are all technologies that provide a great deal of advantages over traditional relational databases with non machine-readable data, why have we not seen more software solutions to support these new data infrastructures? The answer most likely lies in the complexity

of the transition to RDF, especially as the field is still relatively small (in comparison to the larger development community) with few experts who can provide guidance.

Making the commitment to publish Linked Data requires institutions to be cognizant of the options for doing so, together with the respective implications, as those will impact the methodology used for data transformation and publishing. If the switch is one where a legacy system will be maintained, a highly reproducible ETL process will need to be built that will make it easy to take data from the legacy system and transform it to RDF. If no legacy system needs to be maintained, as is the case with the metadata generated from the FotoIndex project, the data transformation process can be performed once, as the master data will live in an RDF-native systems. The transformation process is not however the bottleneck for most research and collections data. The process of cleaning, normalization, and reconciliation as described in the previous chapter is a far more arduous process. If choosing to maintain legacy systems, it is important to be well informed of the limitations of those source systems, as not all legacy documentation systems will allow bulk uploading of data that has been cleaned outside of the system, particularly with commercial solutions where institutions do not have direct access to the underlying database. Additionally, a system may not have the ability to store fields for a particular entity that are not exposed in a public interface. In these cases, reconciliation — adding URI's pointing to external resources — may need to live in separate systems. The alignment of internal entities with URI's to external resources would then need to happen during the ETL process when transforming data to RDF. While maintaining multiple systems to store and present the same data is not optimal, there may not be many other alternatives. Most institutions have a host of workflows, administrative metadata, collection metadata, and even

research data that relate to one another. These data are rarely managed through a single system, despite a desire from most individuals to be able to integrate them. In an ideal world, a single system would be able to cover all of these needs in an RDF native environment.

Data Modeling

Much research has been published on the topic of knowledge representation and data modeling for the cultural heritage domain⁹¹, and the topic has seen a surge of interest as more and more institutions make the transition to RDF. Discussions around the relative merits of certain ontologies (in particular the CIDOC-CRM) have produced a vibrant community of specialists, with many differing perspectives with regard to the way in which these ontologies should be implemented. CIDOC, which is both an ontology and a model, is the oldest and most established knowledge representation system for cultural heritage objects. Managed by the documentation group of the International Council of Museums⁹² (ICOM), the model has been built by multidisciplinary teams from around the world over the last 25 years, even before the advent of Linked Data. It allows for an extremely high level of granularity and expressivity when describing the network of knowledge surrounding cultural heritage. As an event-based model, data is described through a series of events, such as the production or modification of an object. The CIDOC Special Interest Group (SIG), has tried to build the model in an organic and extensible way that can accommodate almost all variables of what might be said about historical

91. Doerr, Martin. "Ontologies for Cultural Heritage." *Handbook on Ontologies*, edited by Steffen Staab and Rudi Studer, Springer Berlin Heidelberg, 2009, pp. 463–86. *Springer Link*, doi:[10.1007/978-3-540-92673-3_21](https://doi.org/10.1007/978-3-540-92673-3_21).

92. ICOM. <https://icom.museum/en/>. Accessed 19 Feb. 2019.

objects, whether they be artworks, architecture, or historical documents. With the ability to grow organically with the increasing complexity of statements, the model can become unwieldy at times when trying to describe a particular object or event. For example, with the historical photographs that came from the FotoIndex project, there are different approaches that one can take when modeling these records when using the CIDOC ontology.

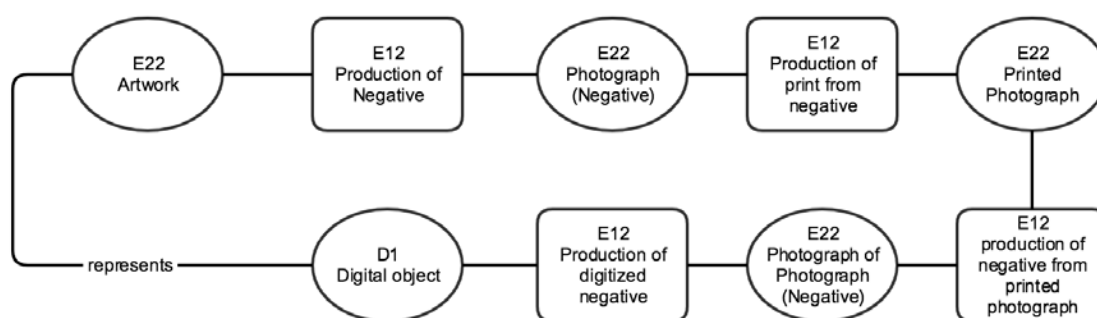


Figure 39: Production of a digital surrogate in the FotoIndex project.

As seen in figure 39, using the CIDOC, we are able to be very expressive about the provenance of our objects. Here we are stating that an artwork was photographed, and a negative was created. From this a print was created, which was subsequently photographed once again, and then digitized. Using production events that act as intermediate nodes allows us to make individual statements about each of those events, the person who took the photograph, which may be different from the person who printed it, or who subsequently rephotographed it, and then digitized it, along with the date and time, location, and methodologies used for each of these events. Statements about these processes allow for a high level of expressiveness that can allow for the description of all of the contextual knowledge around a given event, a feature that is lacking in traditional relational databases where this contextual data is flattened out in favor of

simplicity. The challenge emerges at the application layer when needing to retrieve these data, as a SPARQL query would need to traverse each of those nodes to be able to extract this contextual data. One option to simplify the traversal of these nodes, is simply not to model all of the intermediate events, if one is certain that they do not have any statement that they want to make about their production. Figure 40 depicts a simplified version of the creation of a digital surrogate for the FotoIndex project, where the creation of the original negative and second negative are excluded from the model.

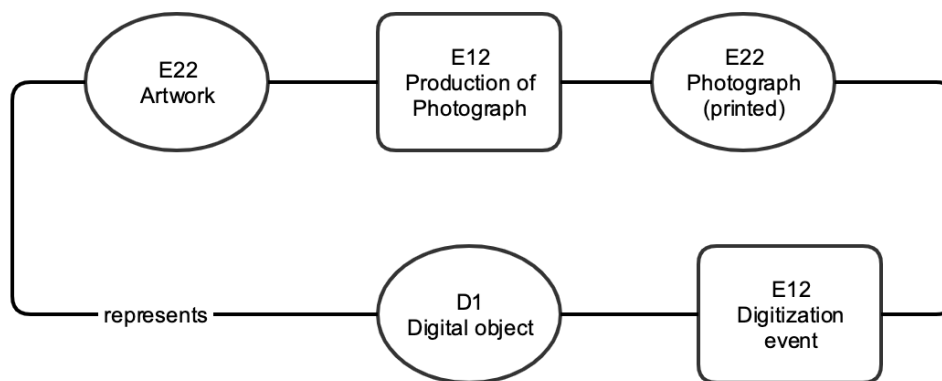


Figure 40: Simplified version of production of a digital surrogate in the FotoIndex project.

This second approach reduces the number of intermediate nodes that one would need to traverse to arrive to the digital object, and would hinder the ability to provide contextual data, such as the creator of the negatives (if different from the printed photograph). This approach facilitates the readability of the data, both when examining the raw RDF data and when building applications that need to extract the labels of these nodes for presentation in a user interface.

Two schools of thought have emerged when it comes to implementing these data models for RDF-based systems. The first approach is to have a shared data model, also known as an “application profile” (such as Linked.art), where common fields in a particular domain are

modeled, described, and mapped in exactly the same way. The benefit here is that when building applications, this greatly facilitates the alignment process at the software layer, since all of the data fits within the same structure and a single query can retrieve all of the relevant records. The downside is that some of the richness of our data is lost, as shown in Figure 40, where many of the intermediate processes to arrive to the digital object are not documented and described. This is the approach favored by the Linked.art movement, where for the purpose of integration, flattening out some of the complexities of cultural heritage data can speed up and facilitate the building of systems with federated search functionality. While creating a shared model can greatly facilitate integration, the approach taken in Figure 39 allows for the full expressivity of these datasets and is the one employed in the FotoIndex project.

Data Integration Methods

Since one of the key principles of Linked Data is to allow for the integration of datasets across the Web of Data, it is critical to employ a data integration methodology that allows for the traversal of the graph using well established standards. If institutions choose to map their data to an application profile similar to Linked.art, it is important to first map the data in a way that captures the full expressiveness of the data, in order to ensure that critical pieces of information are not left out. It is also possible to model the data in more ways than one, having multiple models materialized to the graph. User interfaces can be build around one data model, and a separate model can be used for integration methods. As outlined in chapter two (expressiveness vs. interoperability), the Cultural Heritage domain that deals with publishing data about

artworks, currently has two approaches for integration. Linked.art creates a data layer that uses shared paths for traversing the graph, while FC's and FR's advocate for a methodology that materialize highly generalized categories for entities (persons, places, things etc.) and the relationships between them.

In the FotoIndex project, FC's and FR's were used for a number of reasons. At the time of the creation of the dataset, the Linked.art application profile was not mature enough to provide a stable data model that the author was certain would persist in the immediate future. Additionally, the ability to integrate with data that are not specifically related to artworks, such as historical documents, bibliographic entities, people, institutions, and places, meant that the types of queries that can be performed on the collection become much more conducive to the research practices of art historians. In this way, FC's and FR's allow one to query the complex web of relationships in more organic ways, without being centered on artworks. At the application layer, without materializing FC's and FR's to the data, one would need to build queries (and make them available to users) that traverse all of the variables in intermediate nodes to extract each explicit relationship, which would become unwieldy and could only scale to a certain point with graph databases. By providing a "shortcut" between entities, it can facilitate information retrieval while still allowing datasets to leverage the full expressiveness of the CRM. These categories and relationships are described in full in the journal article *Fundamental Categories and Relationships for intuitive querying CIDOC-CRM based repositories*.⁹³ Finally, this method was chosen over the Linked.art approach as it was already implemented by the British Museum in their ResearchSpace platform.

93. Tzompanaki, Katerina, and Martin Doerr. *Fundamental Categories and Relationships for Intuitive Querying CIDOC-CRM Based Repositories*. 2012, p. 153.

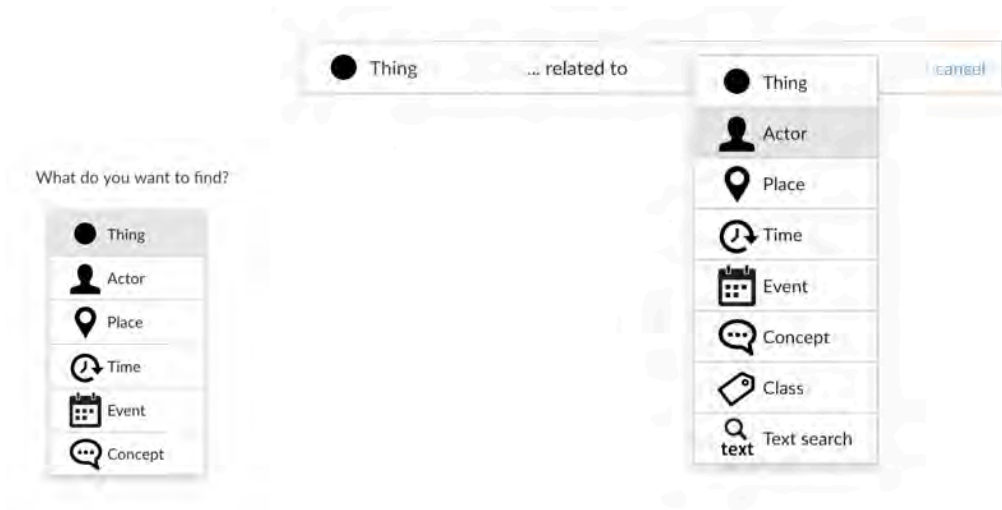


Figure 41: ResearchSpace implementing Fundamental Categories and Relationships

In addition to allowing for integration across datasets, FC's and FR's allow for contextual search functionality in the ResearchSpace platform, as seen in Figure 41. This functionality provides a novel method for exploring data that has hitherto never been seen. Where typical "advanced" search functionality of documentation systems will provide a large set of fields where users can perform textual searches, possibly with the option to add boolean operators, contextual search allows for a far wider array of search parameters. For example, one can search for objects created in a certain region and time, and then explore the distribution of material types used in these objects. Alternatively one can explore the geographic distribution of actors within a specific time and region, allowing for a discovery that is not object-based as is done with traditional documentation systems. This approach offers a paradigm shift in the way that most users experience cultural heritage documentation systems, as it enables new layers of discoverability and analysis, and is one of the principal reasons for which ResearchSpace considers itself a research system rather than a documentation system.

FotoIndex Data Models

A growing number of datasets and collections are actively being integrated at the Harvard Center, each with their own data model and FC and FR alignment. These models have seen multiple iterations over the years, as data modeling is an iterative learning process in constant evolution, especially when the software layer is being developed in parallel. In order to capture the full complexity of data, it was important to involve local experts of the collection with domain knowledge in the modeling process, as they are often the only ones who are aware of the documentation standards that were followed during the original data input. In the case of the collection of images from the FotoIndex project, the metadata was generated with a specific data model already in mind: fields were created with preconceived notions about how they would interact with and be enriched by other (both internal and external) datasets.

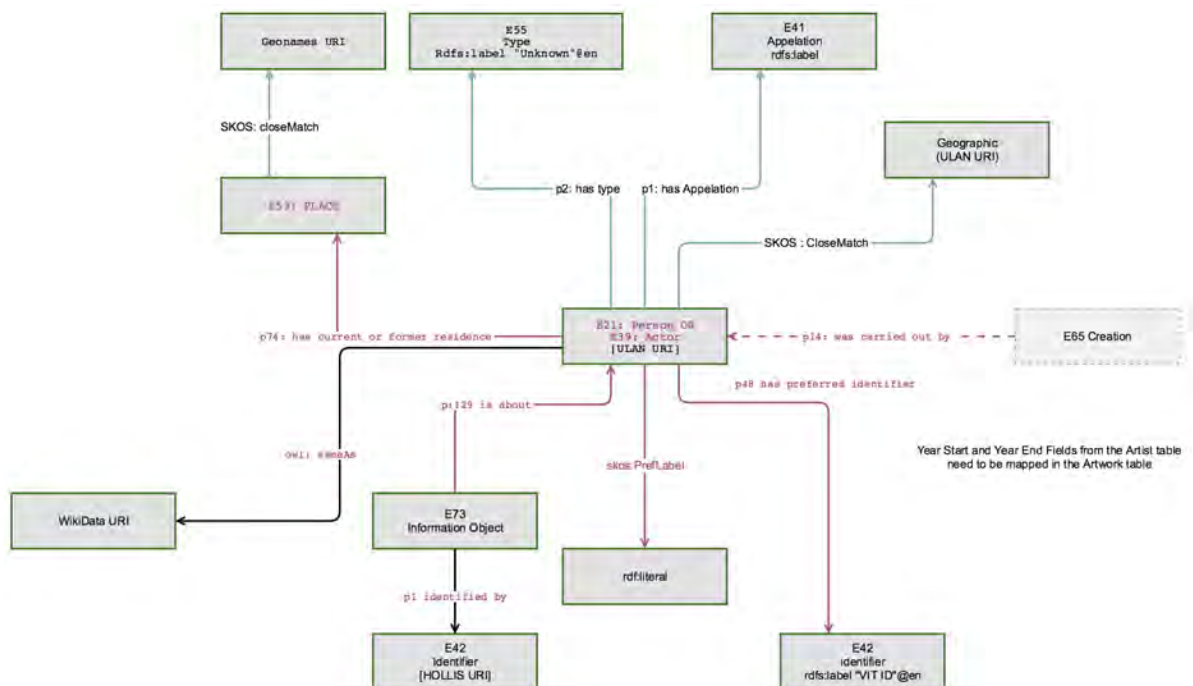


Figure 42: Artist Data model for FotoIndex Project

Data models were created based on available data and centered around specific entities. The artist data model, as seen in figure 42, captures the fields that were generated during the Matchmaker process. Here, the URI of the E21 Person entity is the URI used by the Getty ULAN, or a locally minted URI when the ULAN URI was not available, such as the case with an entity that represents “15th century Florentine” actors. By using external (ULAN) identifiers it becomes easier to align to both external and internal datasets, since those datasets were all reconciled against ULAN and did not use a shared vocabulary from their outset. Reusing URI’s from ULAN allows data from three different projects to be easily integrated: Homeless Paintings, FotoIndex, and Florentine Drawings, together with contextual data from ULAN (date of birth, gender, relationships to other artists, etc). The individual CIDOC modeling choices will not be elaborated on here, as the ontology scope notes serve to provide this documentation and

serve assist institutions in this process. As a general rule, the only other two ontologies that were used were SKOS and OWL to express levels of similarity to other entities. The implementation of these two is the subject of much debate, as the overuse of owl:sameAs has reduced the overall quality of these links across the Semantic Web⁹⁴. In the FotoIndex project, owl:sameAs was implemented only when there was a precise match in the intended real world representation of these two entities, as is the case with a URI in ULAN that represents an artist and the corresponding record in WikiData. In instances where a mere similarity between two entities needed to be expressed, the SKOS:closeMatch predicate was used. This was the case for ambiguous records related to artists from a specific region were being modeled. For a record such as “15th century Florentine”, the concept of “Florentine” was expressed by saying that the individual had a former residence in a given place, and that place was similar to the GeoNames concept for Florence. This is because the concept of the place encompassed by “Florentine” is not the same as the place concept of “Florence” in GeoNames, as historically the geographic and political boundaries of “Florence” have shifted over time.

94. Halpin, Harry, et al. *When Owl:SameAs Isn't the Same: An Analysis of Identity Links on the Semantic Web*. p. 4.

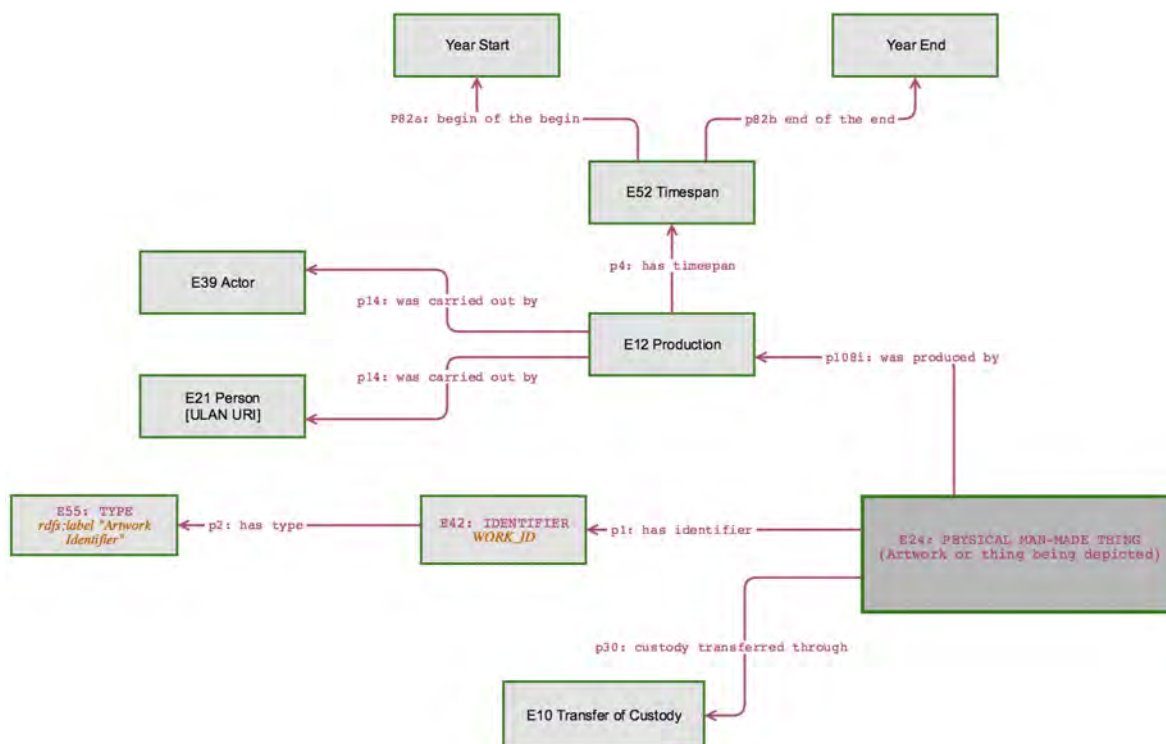


Figure 43: Artwork Data model for FotoIndex Project

The same type of ambiguity was applied to the records of artworks created by these kinds of historical actors. As seen in figure 43, we can see that an E52 timespan with a start date and end date for the actor was materialized. The start date and end date for our “15th century Florentine” record would be 1401 and 1500 respectively. As can be seen by the artwork data model, only a small amount of data was captured in the FotoIndex project about these objects other than the creator and where they were held. Here, the issue of artwork disambiguation will be addressed later on as described in the subsequent chapter, by integrating computer vision functionality to match up artworks to each other and cluster the resulting metadata.

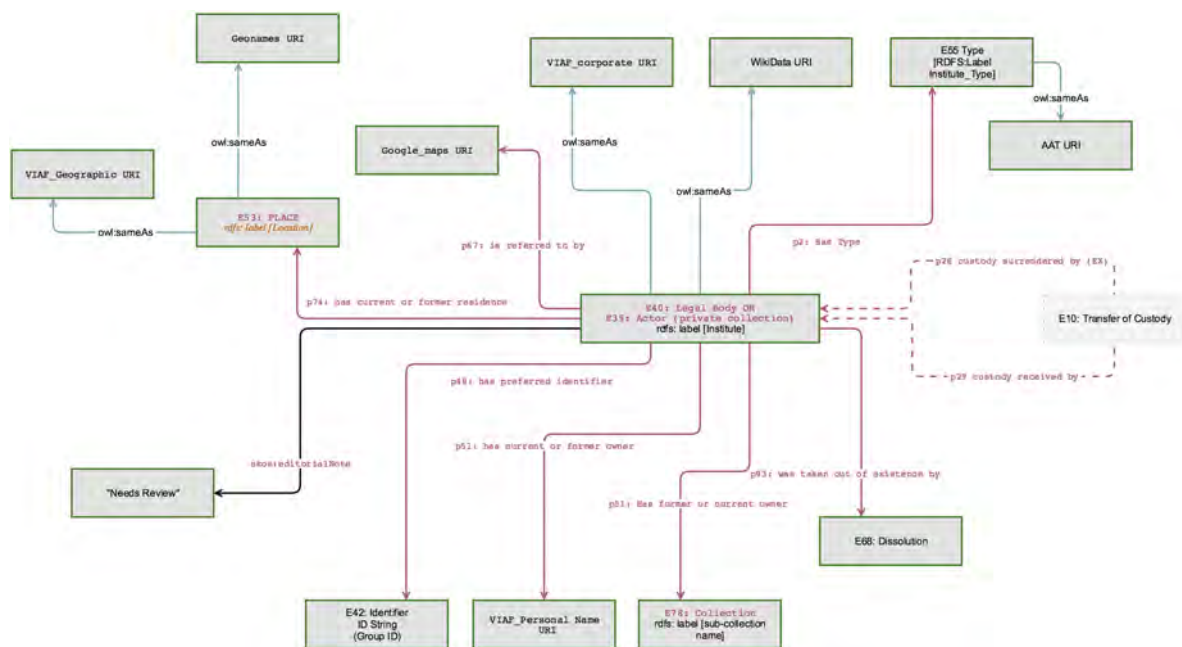


Figure 44: Provenance Data model for FotoIndex Project

The provenance data model (Figure 44) accounts for data captured during the Matchmaker reconciliations process, which was quite rich and expressive. A transfer of custody event allows us to make additional statements about the provenance, specific to that object, such as the name of the institution where it was held, the type of institution, the location, with corresponding URI's to external datasets. A tickbox on the Matchmaker application designated the ownership of the artwork as "Ex" or not, which was represented in the provenance model by making a statement that the ownership had changed since then. Based on the type of institution (private collection or museum or otherwise), conditional statements were made during the transformation process to determine the property that connects this transfer of custody, to the legal body (for an institution) or actor (for private collections).

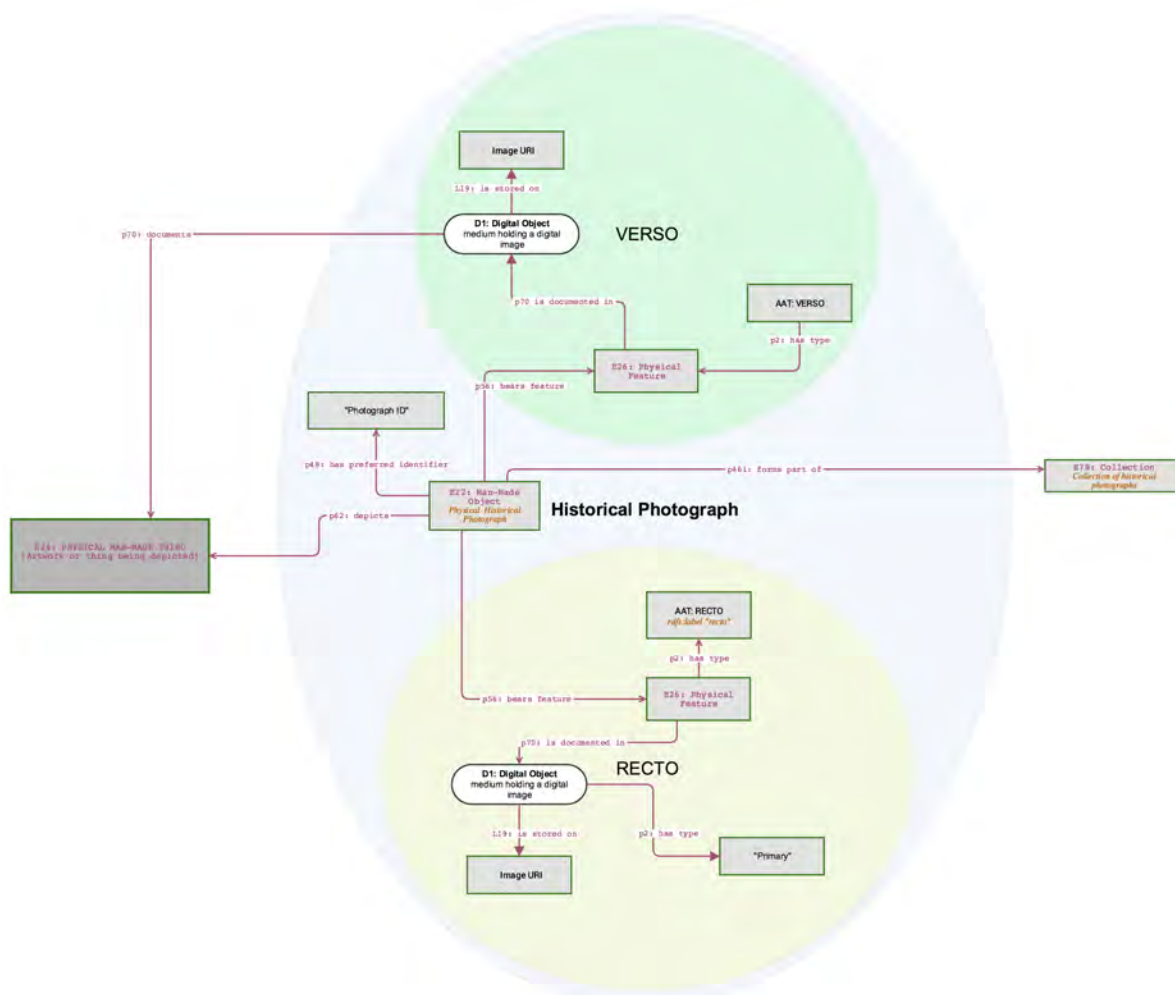


Figure 45: Historical Photograph Data model for Getty Photographic Campaign

As can be seen in Figure 45, the model for the photograph that depicts the artwork omitted data about the real world representation of how the photograph came to existence, so intermediate nodes that represent the photograph negative were omitted as this information was not present in the source data. The historical photograph that documents the artwork is simply listed as having a recto and a verso, each of which has a digital object representing the side. By modeling each side of the photograph explicitly, we are able to make statements about annotations that are written on the back, connecting these statements back to the artwork and ensuring that the provenance of this data is preserved. Preserving the provenance of this data is

especially important over time, since it may not be clear how this information came about, and would therefore not be referenceable by scholars.

Named Graphs

Named Graphs, or quads, can be used in a graph database to group collections of triples with a single identifier⁹⁵. Some lack of clarity in the field regarding the proper use of Named Graphs has caused there to be a wide range of implementations. Proposals have been made that serve different use-cases⁹⁶, such as the tracking the provenance of RDF data, and managing groups of data for replication, versioning, and access control. For example, if one set of collections data makes heavy use of a particular vocabulary (such as ULAN in the case of the FotoIndex project), that dataset can be loaded into the graph database to allow for quicker access to these data, rather than running a SPARQL Service query that dynamically queries the Getty endpoint. The management and versioning of this dataset can then be facilitated by wrapping the entire set of triples in a named graph. ETL processes can be constructed where in order to run an update on that graph from the source, the entire dataset is replaced by the new one, or only the newest data can be ingested and wrapped with a timestamp. Data at any level can also be wrapped in a single graph to be able to denote the provenance. This can be implemented at the

95. *RDF Graph Literals and Named Graphs*. <https://www.w3.org/2009/07/NamedGraph.html>. Accessed 9 Mar. 2019.

96. Dodds, Leigh. "Managing RDF Using Named Graphs." *Lost Boy*, 5 Nov. 2009, <https://blog.ldodds.com/2009/11/05/managing-rdf-using-named-graphs/>.

artwork level, the collection level (such as FotoIndex), or at the level of each triple. CRMInf⁹⁷, one of the many compatible extensions⁹⁸ to CIDOC, is used for “integrating metadata about argumentation and inference”, enabling statements to be made about any other statement. Named graphs can facilitate this kind of argumentation as agreements or disagreements with entities can reference a graph that represents a larger group of triples.

The earlier project undertaken at the Harvard Center on Florentine Drawings used named graphs to delineate different editions of the print publication where the data was derived from. Later it was acknowledged that this was not a proper method for describing the source of these data, as that can be modeled within the dataset itself without adding too much complexity. It is important to note named graphs should not be used for querying data, as the named graph is not returned in a SPARQL query unless it is explicitly requested. Without specifying the graph, all triples will be returned, so they are much more useful when used for administrative metadata. Within the FotoIndex project, named graphs were used to wrap the entire collection of data with an identifier to facilitate the bulk updating of these data continuously, and to be able to separate this collection of data from other collections such as the Homeless painting collection and Florentine Drawings. After much trial and error, this method has been the preferred method that is advocated here, facilitating the administrative management of data from several projects.

97. *ICS - CRMInf: The Argumentation Model*. https://www.ics.forth.gr/isl/index_main.php?l=e&c=713. Accessed 9 Mar. 2019.

98. *Compatible Models & Collaborations | CIDOC CRM*. <http://www.cidoc-crm.org/collaborations>. Accessed 9 Mar. 2019.

Transformation

The process of data transformation involves taking source data (usually in XML format), transforming it to RDF using target ontologies and a chosen data model. There are various existing tools that can assist in the process, or one could build the transformation scripts themselves using any range of programming languages. Karma⁹⁹, a data integration tool developed at the University of Southern California has seen a lot of development over the years, and is capable of accepting a wide range of formats in order to transform them to RDF. Memory Mapping Manger¹⁰⁰ (3M), a tool developed by FORT-ICS, was built specifically to assist in the transformation of data to the CIDOC-CRM and was the tool used to perform data transformation on the FotoIndex dataset.

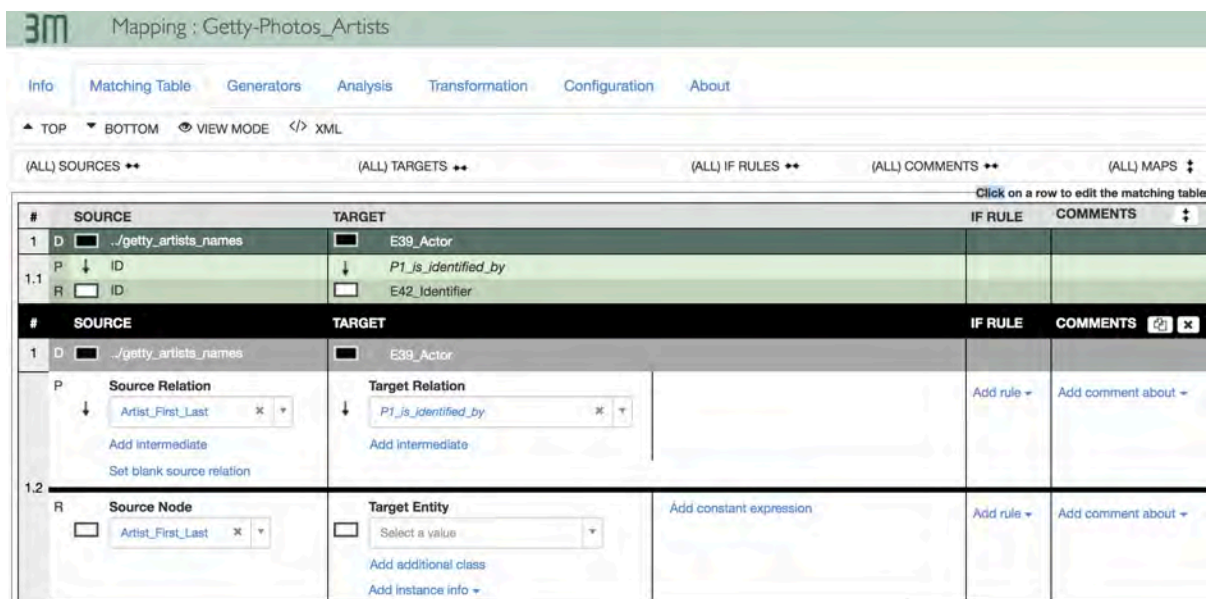


Figure 46: Using Memory Mapping Manager to transform FotoIndex data.

99. Karma: A Data Integration Tool. <http://usc-isi-i2.github.io/karma/>. Accessed 9 Mar. 2019.

100. 3M. <http://www.ics.forth.gr/isl/3M/>. Accessed 9 Mar. 2019.

Sample data is loaded containing all of the possible fields in a data set using a web interface, and the user is then guided through a series of steps to configure the mapping files. Target ontologies are loaded, either in OWL or RDFS format, which provide a framework for mapping as they contain the domain and the range for each target entity, allowing users to select only the corresponding properties for each associated class. Complex mappings can be defined directly in this interface, enabling the inclusion of intermediate nodes that are implicit in the source data (such as a production event). The software can be used on the FORTH-ICS endpoint, or it can be installed locally in a docker container.¹⁰¹ Once the sample data is loaded, and the source data is mapped to the target ontology, a URI generator policy must be created in order to customize the structure of the URI's that are generated. With a prefix "vit" set to "https://collection.itatti.harvard.edu/resource" the following policy would pass in the identifier from the "work ID" field into the {id} variable to generate the full URI:

```
<generator name="workIdentifierURI" prefix="vit">
  <pattern>work/{id}/identifier</pattern>
```

The resulting URI would be:

```
https://collection.itatti.harvard.edu/resource/work/W1000000/identifier
```

It is important that the URI naming strategy creates readable and logical identifiers, and that it can be applied to all of the data that is mapped in the same way. Although there is no single standard for building URI's, the FotoIndex project applied a strategy based on

101. Marketakis, Yannis. *Dockerized Version of 3M. Contribute to Ymark/3M-Docker Development by Creating an Account on GitHub*. 2017. 2019. *GitHub*, <https://github.com/ymark/3M-docker>.

data.gov.uk.¹⁰² This strategy states that real world objects should form the base of the URI, with related concepts building out from there:

```
https://collection.itatti.harvard.edu/resource/work/{identifier}/identifier
https://collection.itatti.harvard.edu/resource/work/{identifier}/{title1, title2}
https://collection.itatti.harvard.edu/resource/work/{identifier}/production/
```

Although there are differing schools of thought on the use of labels within a URI (such as the title of an artwork), the FotoIndex project followed the URI generation strategy employed by most other larger institutions such as the British Museum and WikiData. While adding the label may provide more readability:

```
https://collection.itatti.harvard.edu/resource/work/W1000000/title/Le_gôûter
```

One can see how this can quickly become problematic with the necessity to implement escape characters for spaces and diacritics, along with a lack of consistency with various interpretations of titles or different languages. URI's for labels were instead generated using a simple numbering policy when there was more than one, e.g. {title1} or {title2}. Once the process of mapping the source data to the target ontology is complete, the 3M web interface will allow the user to check the sample data output and export a mapping file. These configuration files are used to pass into the x3ml engine¹⁰³ which needs to be run locally as a java applet. The resulting output is an RDF file that is ready for ingestion into a graph database or triplestore.

102. *Creating URIs* | *Data.Gov.Uk*. 15 July 2017, <https://web.archive.org/web/20170715074122/https://data.gov.uk/resources/uris>.

103. <https://github.com/isl/x3ml>

Publish

The process of publishing Linked Data has been greatly facilitated by platforms such as ResearchSpace and Metaphactory, as they allow for the loading of RDF datasets, then constructing user interfaces around these data. When the FotoIndex project was initially conceived by the author in 2014, these systems were not available and the intent was to use libraries, such as `rdflib.js`¹⁰⁴ that allow for the reading and writing of RDF datasets to publish the user interface. At the time, RDF was being used as a data format that allowed for the sharing of raw data, more than serving as the basis of a web platform. In 2017 with the digital edition *The Drawings of the Florentine Painters*, the author began using the ResearchSpace system to store and publish the raw data, but it was not yet mature enough to allow for a user interface that enabled search and discovery. Since then, the platform has evolved with a wide range of features, very well suited to the cultural heritage domain, with a data-centric architecture making it a very attractive option for publishing these data.

ResearchSpace is an open-source alternative based on the commercial product Metaphactory by the company Metaphacts, and is an end-to-end platform that provides the full suite of tools necessary to be able to load up RDF data, build interfaces that allow you to visualize, search, author, and manage your records. Content Management System (CMS) functionality enables the development of new plugins and modules that can make use of the RDF data through SPARQL queries. Existing tools include geographic, time, and chart visualizations, IIIF image annotation, text document annotation, a clipboard for users to be able to save and

104. *Linked Data API for JavaScript. Contribute to Linkeddata/Rdflib.js Development by Creating an Account on GitHub*. 2011. Read-Write Linked Data, 2019. *GitHub*, <https://github.com/linkedata/rdflib.js>.

recall searches or entities, among others. As an end-to-end platform, the Harvard Center has decided to do a full migration to ResearchSpace for use as a documentation management system, as authoring functionality allows you to edit records with relative simplicity, and linking to external datasets can be done on-the-fly.

The primary functionality revolves around a powerful templating engine, where pages, or “records”, are constructed that combine HTML5 elements with SPARQL queries to populate collection data on the page. Based on the URI of the entity that a user wants to visualize, a templated is called that retrieves that particular record. For example an “E39 Actor” entity has a type which retrieves data that is relevant to that entity.

```

1- <style>
2- .metaphacts-carousel-widget{
3   overflow: visible;
4   position: relative;
5   width: 220px;
6 }
7 /* to hide the back- and forth carousel buttons if there are no further images */
8- .slick-disabled{ display: none !important; }
9 </style>
10
11- <bs-row>
12-   <bs-col sm=8 sm-offset=2 >
13-     <ol class="breadcrumb" style="background:white;border:none;padding-left:0px;">
14-       <li>
15-         <a title="Portal" href="/">Home</a>
16-       </li>
17-       <li>
18-         <semantic-link title="Drawings" uri="http://vocab.getty.edu/aat/300033973">
19-           Drawings
20-         </semantic-link>
21-       </li>
22-       <li class="active">[[this.label]]</li>
23-     </ol>
24-     <!-- header row -->
25-   <bs-row>
26-     <bs-col sm=4>
27-       <div style="float:left;">
28-         <semantic-carousel query="SELECT DISTINCT ?image WHERE{
29-           OPTIONAL{? ? crm:P46_is_composed_of/crm:P1381_has_representation/crm:P1381_has_representation ?img. }
30-           BIND(COALESC(?img, "https://upload.wikimedia.org/wikipedia/commons/thumb/a/ac/No_image_available.svg/200px-No_image_available.svg.png") as ?image)
31-         }"
32-         layout="{
33-           'tupleTemplate': '<mp-overlay-dialog title='\<Image Detail\>' bs-size='\<large\>' type='\<modal\>'>
34-             <mp-overlay-dialog-trigger><img src='\<{image.value}\>' style='\<max-width: 200px; max-height: 200px; cursor:pointer;display:block;margin:0
35-             auto;\></mp-overlay-dialog-trigger>
36-             <mp-overlay-dialog-content><div class='\<text-center\>' ><img src='\<{image.value}\>'></div></mp-overlay-dialog-content>
37-           '</mp-overlay-dialog>',
38-           'options': {'infinite':false, 'variableWidth': false}

```

Figure 47: Using the ResearchSpace to build record templates.

As seen in Figure 47, templates can easily be constructed and modified in a graphical user interface. HTML5 is used to configure the placement of various fields retrieved via

SPARQL, and a series of included components allow for relatively simple front-end development in the user interface, as shown in Figure 48 which demonstrates how that template is translated to a record view.

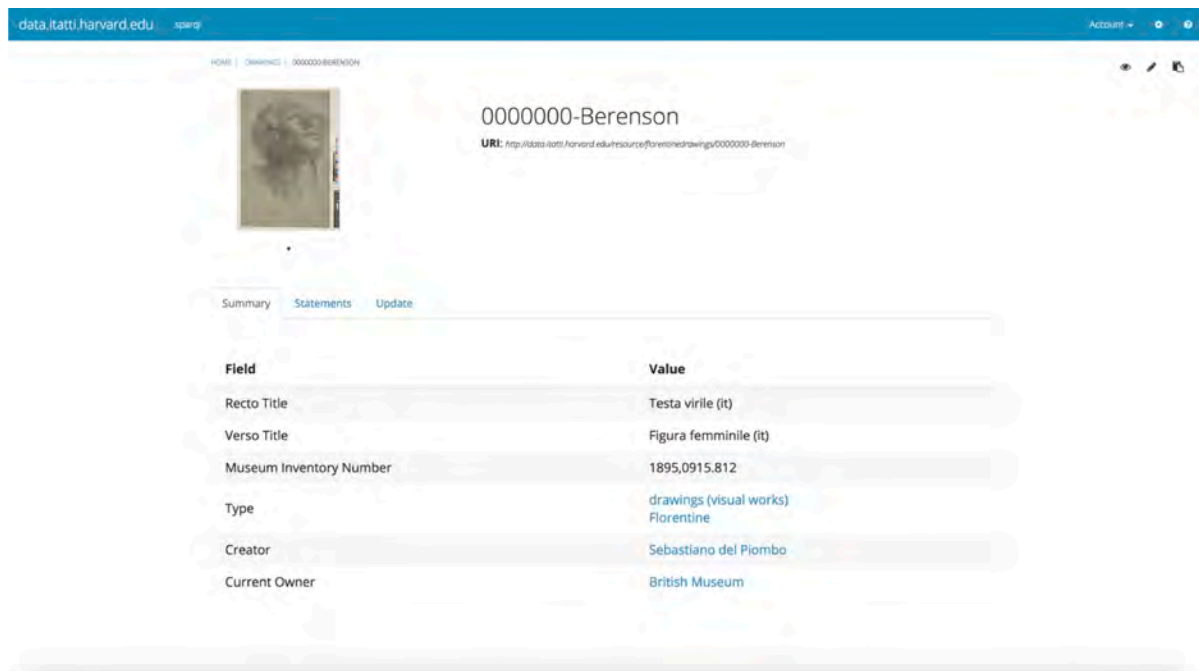


Figure 48: Sample view for a record

Another component that makes the architecture of this product extremely versatile, is a field definition (or “knowledge pattern”) feature that allows you to use SPARQL queries to map a pattern in your RDF data to a certain field name, which can then be called programmatically and reused across the platform in the visualization, search, filtering, and the authoring of records.

The screenshot displays the 'Field: DrawingTitle' configuration page in the metaphactory application. The interface includes a sidebar with various field configuration options and a main content area for editing the field's properties.

Field Properties:

- Label:** DrawingTitle
- Identifier:** `http://www.metaphacts.com/instances/fields/DrawingTitle`
- Description:** (empty)
- Categories:** (empty)
- Domain:** (empty)
- XSD Datatype:** `xstring`
- Range:** (empty)
- Min. Cardinality:** (empty)
- Max. Cardinality:** (empty)
- Default values:** (empty)
- Test Subject:** (empty)
- Insert Pattern:**

```
1 - INSERT {
2   ?title rdf:type label ?value.
3 - } WHERE {
4   $subject crm:P102_has_title ?title.
5 }
```
- Select Pattern:**

```
1 - SELECT ?value WHERE {
2   $subject crm:P102_has_title/rdfs:label ?value.
3 }
```
- Delete Pattern:**

```
1 - DELETE {
2   ?title rdf:type label ?value.
3 - } WHERE {
4   $subject crm:P102_has_title ?title.
5 }
```
- ASK Validation Pattern:** (empty)
- Value Set Pattern:** (empty)
- Autosuggestion Pattern:** (empty)
- Tree Patterns:** (empty)

Field Preview: Shows a plain text input field with the placeholder text 'Enter drawingtitle here...' and a '+ Add drawingtitle' button.

JSON Configuration Example:

```
[{"selectPattern": "SELECT ?value WHERE {\\n $subject crm:P102_has_title/rdfs:label ?value.\\n}", "xsdDatatype": "http://www.w3.org/2001/XMLSchema#langString", "defaultValues": [], "id": "example", "label": "DrawingTitle", "insertPattern": "INSERT { \\n ?title rdfs:label $value.\\n } WHERE {\\n $subject crm:P102_has_title ?title.\\n}"}]
```

Figure 49: Field definition implementation

In figure 49, we can see the portion of our dataset that represents the Drawing Title, which has a specific pattern in our dataset. By mapping these data to a more user-friendly field,

the management of data is greatly streamlined. These field definitions can also be used to materialize FC's and FR's in a dataset during authoring.

Linked Data Platform

Publishing Linked Data on the web is in practice, still a field that is being tested and defined. While the original vision of the semantic web envisaged a web of data where information flows freely across institutional silos, the practicalities of implementing this kind of architecture has been hindered by the lack of software that can leverage much of this functionality. The World Wide Web Consortium approved a series of recommendations in 2015 outlining a set of rules defining the architecture of read-write Linked Data applications¹⁰⁵, under a new standard called Linked Data Platform (LDP). These recommendations provide a framework and a broad set of standards that have the capacity to alleviate many of the problems we are faced with when thinking about the architectural framework for Linked Data applications. They were an important step in moving towards RDF-native systems, away from implementations where we keep legacy systems and attempt to add layers of LOD functionality. As the recommendation is still relatively new, not many systems conform to these standards, especially in the open-source sphere. CabonLDP¹⁰⁶ and Callimachus¹⁰⁷ are other products that also act as a middleware between the graph database and the user, both with free licenses, but

105. *Linked Data Platform 1.0*. <https://www.w3.org/TR/ldp/#ldpc>. Accessed 10 Mar. 2019.

106. *Home | Carbon LDP*. <https://carbonldp.com/>. Accessed 10 Mar. 2019.

107. *Callimachus - Data-Driven Applications Made Easy*. <http://callimachusproject.org/>. Accessed 10 Mar. 2019.

neither are mature enough or able to provide the full suite of tools necessary to manage research and collections data. One very recent project that does show a great deal of promise, is the Solid project¹⁰⁸ spearheaded by Tim Burners-Lee, the inventor of the World Wide Web. Launched in September of 2018, the project promises a new web infrastructure which “rethinks how web apps store and share personal data”¹⁰⁹. The project is still in its early stages of development, so its outcome is still unknown, but it does aim to tackle many of the same issues as the LDP architecture.

Shifting to RDF-native solutions, while ideal from the perspective of data management and interoperability, still have scalability issues as graph databases cannot yet scale to the levels of other storage engines. The cultural heritage domain, however, has relatively meager amounts of data and scalability should not be an issue at this time. Solutions such as those built by Metaphacts and CarbonLDP seem to be the most reasonable approach: a middleware acts as a layer in between the graph database/tripplestore to query and visualize the data, but also allows for programmatic querying from other systems.

Architectural Issues

RDF-native infrastructures are faced with a few architectural challenges that other data stores do not grapple with. These are slowly being overcome in the commercial sector, but these underlying architectural limitations bring additional overhead at the application layer. Unlike

108. *Solid | Inrupt*. <https://inrupt.com/solid>. Accessed 10 Mar. 2019.

109. Orphanides, K. G. “How Tim Burners-Lee’s Inrupt Project Plans to Fix the Web.” *Wired UK*, Feb. 2019. www.wired.co.uk, <https://www.wired.co.uk/article/inrupt-tim-berners-lee>.

SQL databases, graph stores and other noSQL storage solutions make the managing of permissions more challenging, both at the read and write level. Whereas applications built on top of relational databases are able to apply varying levels of permissions to tables and functional features of the database, the architecture of graph databases tend to offer an all-or-nothing approach. Blazegraph, the graph database usually implemented in the backend of ResearchSpace, has no users or permissions of any kind. The same is the case for Amazon Neptune and many other graph stores available on the market. Some solutions are able to provide some kind of permissions functionality, but the underlying architecture of graph databases does not lend itself well to fine-grained access controls on data, as the data is lacking the structure necessary to implement these access controls. The expectation is that this functionality will be handled at the application layer. With a system where one of the core features is the flexibility of creating new data that does not conform to preexisting data structures, it becomes problematic when deciding who can access which data and at what level. ResearchSpace implements permissions management in two ways that does not satisfy all use cases and may cause the system to be either too open or too closed to certain users. Basic permissions management allows users to either read or write anywhere. Another solution implements multiple graph databases, where access is managed based on the repository. The most granular level of access allows all data access to pass through a REST API, where every request to read or write data is parsed through the API that translates the request into a SPARQL query. While this provides fine-tuned access to the data it does not allow for the more organic building of research data, as every SPARQL pattern must be pre-configured and translated to an API call, with corresponding permissions levels. For this

reason, these kinds of systems are generally not well suited for administrative metadata— such as who consulted an object, if it was on loan at a certain time, etc.

Another underlying architectural issue with Graph databases is the lack of a robust search backend. SPARQL does allow for textual searching through regular expressions, but the performance and reliability of this solution is not ideal. Most graph databases have implemented another search backend, such as an Apache Lucene or Apache Solr index. The database backends create a full-text index of the entire content of the database, and exposes the content through a special predicate that can be used in the SPARQL query (bds:search in the case of Blazegraph). While this does improve speed and efficiency, relevancy ranking is problematic to configure given the vast number of entity types and more dynamic database structure.

Impact and Future Work

The outcome of the FotoIndex project is multifarious: share methodologies associated with the mass-digitization of the collection, providing a framework for other institutions to be able embark on large digitization projects without the prospect of having staff that can create a full inventory of these records. Provide a toolset (Matchmaker) and document methodologies for the cleanup, reconciliation, and enrichment of collections data. Provide guidance on the implementation of an infrastructure that can make full use of these semantically enriched data, and finally publish the dataset openly so that other individuals and institutions can explore, interpret, and create links of their own.

At the outset of the project, it was anticipated that this data would be published on a collections portal for the research institute, together with other collections that are owned by the Harvard Center. While this is still the case, other projects have grown out of this initiative that are more focused on research data, rather than collections data, where scholars are able to annotate and transcribe historical documents and publish new scholarship in the form of articles. While research data can augment collections data, it is also important to make a distinction between them. Therefore data from the FotoIndex project, together with the Homeless Paintings project, the institutional art collection, and the archive will be published as static resources on collection.itatti.harvard.edu, using that URI as the stable identifier. That platform will also provide cataloging functionality for cataloging staff to be able to modify and augment those records. Research projects, such as *The Drawings of the Florentine Painters* digital edition will be published on a research portal that will provide different views of the data, and allow the scholarly community to comment, make new assertions, and publish semantically enriched scholarly articles on the platform. The domain name ArtResearch.net has been secured for this purpose and will provide a wide range of functionality to support the research practices of scholars in the community.

Providing functionality that enables the continuous enrichment of data through a suite of tools that can constantly be improved and adapted to suit various research projects is the most sustainable approach to being able to manage research data and projects long-term. Rather than creating silos of data on various websites, we can create building blocks for projects that can be reused and shared for different implementations. Every year at the Harvard Center, a growing number of research fellows have digital projects and data that they would like to publish. The

implementation of this research environment will address the challenge of providing a platform where non-technical users can upload and produce research data. The overarching goal is to provide generic publishing services that allow for the tight integration of scholarly articles with semantically enriched research data, visualizations and annotations, building a community of scholarship in a research infrastructure that is easily reproducible and sustainable.

V - Visual Search for the Semantic Web

The fields of Computer Vision (CV) and Machine Learning (ML) have seen substantial progress in recent years, with the majority of technology titans (Google, Amazon, Microsoft, etc.) recognizing their usefulness and applicability to a wide range of domains. Companies have created their own tools and exposed them through API's available with a pay-per-use model, and open-source tools have flooded the web, with individuals and institutions exploring the application of these technologies to ask broad questions to large collections of images. The field of Digital Art History stands to benefit a great deal from these tools, as numerous projects have already made good progress in tackling issues of visual similarity, artwork classification, style detection, and gesture analysis, among others. While some attempts have been made to create more generic tools that could be used across a wide range of domains, most tools are lacking in reusability beyond the specific use-case for which they were built. This chapter argues that there is no one-size-fits-all toolset that can address the wide array of needs required for the field of Digital Art History. For example, models that can be used to find other images of the same artwork, employ methodologies that are completely different from those that serve to find visually similar artworks. While this chapter provides a brief, non-comprehensive overview of some of the toolsets that are available for performing visual searches on artworks, it primarily seeks to propose and advocate for a semantic framework and system architecture that allows for the integration of multiple CV and ML models within a single web platform. This framework is extensible enough to accommodate for various models, while being sufficiently expressive

semantically and computationally actionable. The service will most importantly provide computer vision-based artwork disambiguation functionality, enabling institutions to parse images of two-dimensional artworks through a SPARQL endpoint, having those images matched to those of other institutions.

Introduction

Computer Vision for cultural heritage, in particular when applied to two-dimensional artworks, has been a topic of growing interest to scholars and institutions. Various conferences, symposia, and workshops have been organized over the years to explore the usefulness of this technology, generally without any particular tool emerging as a forerunner¹¹⁰. While these tools have made much more headway in commercial sectors, the cultural heritage domain is still behind in this frontier. This chapter takes a look at the landscape of available tools and their applicability to cultural heritage, proposing a solution that aggregates the results of various CV services semantically through a single endpoint. At the most basic level, this framework can provide artwork disambiguation services to individuals and institutions seeking to match up images of artworks across the web. A graph database serves as the backbone of this service, and exposes image similarity functionality through a SPARQL endpoint, using an extensible data model that is able to express an array of image similarity results semantically. Published as a web service, various CV services can be exposed through a single endpoint, with the objective to

110. See: *VISART*. <https://visarts.eu/>
Searching Through Seeing: Optimizing Computer Vision Technology for the Arts | The Frick Collection. https://www.frick.org/interact/video/searching_seeing,

democratize the accessibility of this functionality to non-technical users. Within this framework, researchers could be allowed to build their own models specific to their research needs, in order to classify and be able to search for any kind of visual feature, by training a model on sets of images.

While many vocabularies are available to disambiguate and provide identifiers for artists (ULAN), places (TGN, GeoNames), Institutions (VIAF), and bibliographic entities (OCLC) in the Linked Data sphere, the challenge of providing identifiers to artworks has never been properly tackled. The Getty CONA vocabulary¹¹¹ has made some efforts in this front, but issues of artwork ownership and authority have overshadowed these efforts. This framework seeks to overcome these challenges by serving as a linking mechanism between institutions that are owners of artworks, institutions that have reproductions of these artworks, and scholars who seek to reference them through Linked Data services.

At the time of writing, some commercial and open-source tools publicly available were evaluated, provided that those tools allowed collections of images to be uploaded and parsed through an API. For the purpose of this evaluation, historical photographs documenting Early Modern paintings from the FotoIndex and Homeless Painting collections were used for testing. Commercial services such as Clarifai, IBM Watson, Google Cloud Vision, Amazon Rekognition, Microsoft Computer Vision API, Cloustron, as well as open source products such as Pastec, Pavlov Match, and Tensorflow models were used for the basis of this study. In most cases, these tools were used without specific customizations being developed. Other tools, such as the *Replica* project¹¹² (now published under the timemachine.eu initiative) developed at the DH lab

111. *Cultural Objects Name Authority* (Getty Research Institute). <http://www.getty.edu/research/tools/vocabularies/cona/>. Accessed 10 Mar. 2019.

112. *Diamond*. <https://diamond.timemachine.eu/>. Accessed 10 Mar. 2019.

of the Ecole Polytechnic Federal du Lausanne (EPFL), or those developed for commercial purposes such as ArtPi¹¹³, were not evaluated as they were not available publicly or did not offer any documentation.

Background

Art Historians have a long tradition of writing about images, one could say, transcribing a visual language to a textual one. As with all translation activities, there is a loss of meaning in this process. While the perception of images is instantaneous and universal, writing is bound by time and cultural bias.¹¹⁴ Computer vision has emerged as a powerful tool for the field of Digital Art History with the potential to bridge some of this gap between images and text, allowing images to dialog with one another without the mediation or dependence on text. A great deal of experimentation has already been done in the field: attempts to automatically classify paintings,¹¹⁵ recognize style,¹¹⁶ object detection studies to identify objects in artworks,¹¹⁷ evaluating influence¹¹⁸, matching up images across collections and comparing metadata¹¹⁹, and

113. *ArtPi - Artrendex*. <http://www.artrendex.com/artpi>. Accessed 10 Mar. 2019.

114. Elkins, et al. *What Is an Image?* Pennsylvania State University Press, 2011.

115. Tan, W. R., et al. "Ceci n'est Pas Une Pipe: A Deep Convolutional Network for Fine-Art Paintings Classification." *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 3703–07. *IEEE Xplore*, doi:[10.1109/ICIP.2016.7533051](https://doi.org/10.1109/ICIP.2016.7533051).

116. Karayev, Sergey, et al. "Recognizing Image Style." *ArXiv:1311.3715 [Cs]*, Nov. 2013. *arXiv.org*, <http://arxiv.org/abs/1311.3715>.

117. Crowley, Elliot J., and Andrew Zisserman. "In Search of Art." *Computer Vision - ECCV 2014 Workshops*, edited by Lourdes Agapito et al., Springer International Publishing, 2015, pp. 54–70.

118. Elgammal, Ahmed, and Babak Saleh. "Quantifying Creativity in Art Networks." *ArXiv:1506.00711 [Cs]*, June 2015. *arXiv.org*, <http://arxiv.org/abs/1506.00711>.

119. *John Resig - Building an Art History Database Using Computer Vision*. <https://johnresig.com/blog/building-art-history-database-computer-vision/>. Accessed 10 Mar. 2019.

large-scale analysis of broad concepts within artworks such as gesture.¹²⁰ The opportunities offered by this technology are numerous but the solutions are generally ad-hoc, lack broad applicability, and most importantly the raw data that is produced from this research is rarely published. It is also important to note that the analytical capacity of CV technology generally does not surpass that of an art historian, but is simply able to perform very specific tasks at a scale that would otherwise be unattainable by humans, so it is important to remain cognizant of this limitation and try to build to its strengths rather than weaknesses.

For institutions looking to enrich digital collections of images that document artworks, performing a large-scale analysis with computer vision tools can be a particularly attractive approach to solving a series of issues that were traditionally dealt with manually. For scholars, the opportunity to perform visual similarity searches, or search for visual features within artworks, can open up new doors in their scholarship. While a detailed description of the history of computer vision is not within the scope here, some of these use-cases will be examined that may be applicable for the researcher and institutions looking to looking to enrich their image collections with metadata and provide new access points for discovering images.

120. Impett, Leonardo, and Sabine Süssstrunk. "Pose and Pathosformel in Aby Warburg's Bilderatlas." *Computer Vision – ECCV 2016 Workshops*, edited by Gang Hua and Hervé Jégou, Springer International Publishing, 2016, pp. 888–902.

Visual Similarity

Visual similarity search is perhaps the most basic but also one of the most desirable functionalities for researchers and institutions. Leaving aside the various notions and discourses on similarity, the principal difference between the available tools is the methodology used. Broadly speaking, a “fingerprint” for each image is created, which is then compared against others to identify their “closeness”. In the case of Convolutional Neural Networks (CNNs), this fingerprint is in the form of a vector (see Figure 50), and is created with a pre-trained model that measures a series of “features” for each image.

```
{ "vector":
  [-13.865196580949014, 2.9991287374111035, -2.6954291082404023, 2.523
  5974397768275, -1.1855909562072457, -2.152675725976077, 3.1140508686
  06226, -0.3347796156821097, -3.5947896259184082, -3.8622545414605507
  ,
  2.031925457383388, 0.6457769882400669, -1.1023191207691978, 2.054630
  077528512, -0.9247706215821905, 0.023765016311060733, -1.09754226039
  8586, 1.3144235836283564, -1.1900126390430366, -0.27685082158892194]
}
```

Figure 50: Image Similarity Vector

A measurement is calculated for each feature, and “similarity” is measured by calculating the distance between the numbers of different vectors. The images with which these neural networks have been trained are instrumental in how they will perform for a given task.

Another method for measuring similarity is the bag-of-words methodology¹²¹. By converting parts of an image to a list of “words”, this method has proven to be very successful in

121. Yang, Jun, et al. “Evaluating Bag-of-Visual-Words Representations in Scene Classification.” *Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval*, ACM, 2007, pp. 197–206. *ACM Digital Library*, doi:[10.1145/1290082.1290111](https://doi.org/10.1145/1290082.1290111).

classifying images according to the objects they contain, and remains unaffected by the position or orientation of objects in the image.

Tool Comparison

For the purpose of implementing visual search within the FotoIndex collection, various tools have been evaluated to test their usefulness on this collection. The matrix in figure 51 provides an overview of the services and functionalities that were evaluated:

	exact image match	visually similar	partial image match	image labels/tags	custom model building
Google Cloud Vision		x ¹²²		x	
Amazon Rekognition				x	
Clarifai	x	x		x	x
Pastec.io	x	x	x		
Pavlov Match	x				
Tensorflow Inception V3	x	x		x	x
IBM Watson				x	x
Microsoft Computer Vision API				x	x
Cloudsight				x	

Figure 51: Matrix of Computer Vision tools and their functionality

This list is by no means comprehensive for the field, and does not include tools that were not openly accessible (free or commercially), or their implementation was too cumbersome and not mature enough to use at a larger scale. Other tools that are available that have not been tested

122. note: Google Cloud Vision does not allow visual image search within a specific collection, rather it searches the web for visually similar images.

include Visual Search by Machine Box¹²³, and Deep Video Analytics¹²⁴ by Akshay Bhat from Cornell University, as well as the Bing Visual Search API¹²⁵.

Image Tagging

Parsing a set of historical photographs against six image tagging tools made it evident that using generic models made available by commercial services was not useful in obtaining useful image tags. While these tagging services may be effective in tagging photographs of dogs and balloons (the datasets that they were often trained on), tests on historical images of artworks brought forward their limited usefulness for this use-case.

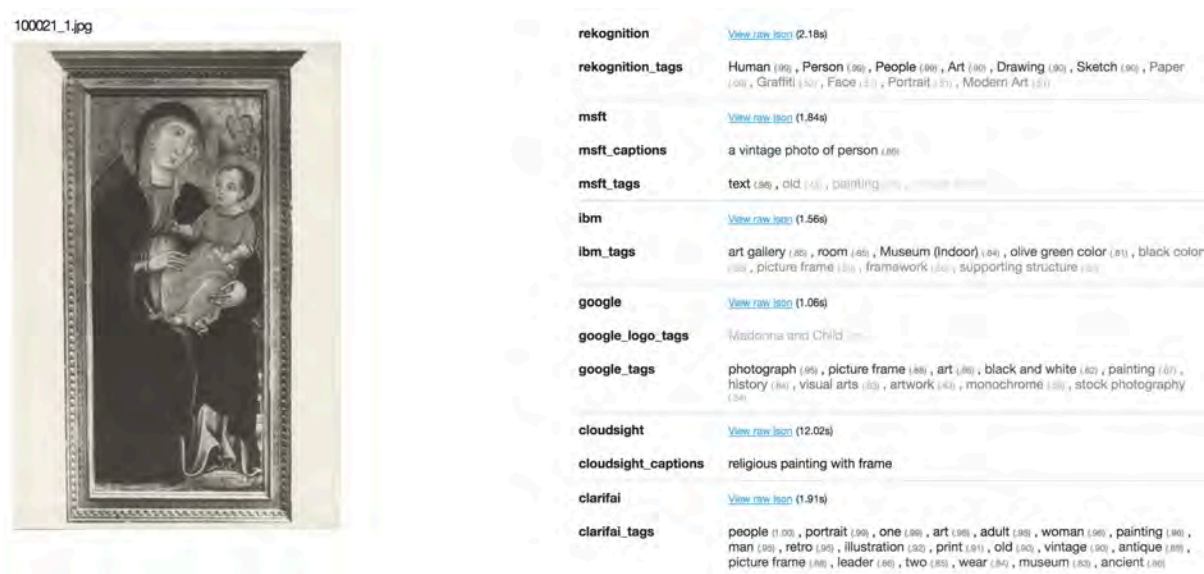


Figure 52: results from six image tagging services

123. Hernandez, David. "Visual Search by Machine Box." *Machine Box*, 7 Oct. 2017, <https://blog.machinebox.io/visual-search-by-machine-box-eb30062d8abe>.

124. *Deep Video Analytics*. <https://www.deepvideoanalytics.com/>. Accessed 10 Mar. 2019.

125. *Bing Visual Search Developer Platform*. <https://www.bingvisualesearch.com/docs>. Accessed 10 Mar. 2019.

As seen in figure 52, the tags that were provided would not be able to improve any kind of access to the image. The generated tags, such as “human”, “photograph”, and “art”, would not provide any additional points of access for a text-based search within these collections. One minor exception was with results from the Google Logo Tags service, that allows one to search for images of logos across the web. In this case, Google recognized part of the image from another image on the web, and mistook it for a logo. Since the other image on the web was tagged as “Madonna and Child”, the same result was returned for this image.



Figure 53: Google Logo tagging service results

While these results may be marginally useful for applying a broad set of labels to images, a very low percentage of images returned any results with the Google Logo Tag service. Additionally, this kind of simple classification could be done on large sets of images much more effectively, using a generic visual similarity search that allows the subsequent tagging of these images. Given the large number of “Madonna and Child” images within the collection, a custom model could be built to classify these iconographical themes in a way that would be much more effective. The full results of these tests, as published on Github¹²⁶, demonstrate that generic commercial tagging services will not currently improve accessibility to historical photographs of artworks, and at best may marginally augment the usefulness of other tools when used in combination. Alternatively, custom-built tagging models could be built that are specific to artworks, but the amount of work necessary to provide an effective image tagging service that can serve a broad range of artworks is quite high.

Most of the ad-hoc work on applying labels to images of artworks has been done by training the last layer of a model in a CNN, using large collections of pre-tagged images as training data. Since these tools are largely the result of individual and institutional research projects, the author is not aware of any of these tools being made publicly on the web or through an API that would allow for integration into a Semantic Web application that brings the results of various tools together.

126. *Cloudy Vision, a Comparison of Image Tagging from Various Vendors for the VIT Photo Archive Collection*. https://lklic.github.io/compare_vision/output/output.html. Accessed 10 Mar. 2019.

Visual Search

As outlined in the matrix in figure 51, there are three forms of visual search that were evaluated, each with distinct use-cases for both institutions and scholars: exact image match, visually similar, and partial image match.

An exact image match, meaning that two images are nearly identical, is computationally and programmatically a much more trivial a task than visually similar images or those that allow for partial image matching. In this case, it is assumed that the crop and content of images would be nearly identical, with only minor variations. Many methodologies can calculate this kind of similarity, but the implementation that was the most functional for this use case was Match¹²⁷, a reverse image search based on ascribe/image-match.¹²⁸ This implementation provides a simple signature for each image which is stored in Elasticsearch for quick retrieval. It scales very well to billions of images, a functionality that few other available products can claim. Although the use case for a near-exact image search is limited, given its limited infrastructure requirement, it could be an excellent tool to implement alongside a service that provides more “fuzzy” similarity searching.

Searches for images that have varying degrees of “similarity” is not as trivial, and the effectiveness of the search has many varying factors. Earlier tests by John Resig¹²⁹ on behalf of

127. *Crystal_ball: Scalable Reverse Image Search Built on Kubernetes and Elasticsearch: Dsys/Match*. 2016. Distributed Systems, 2019. *GitHub*, <https://github.com/dsys/match>.

128. *EdjoLabs/Image-Match: Quickly Search over Billions of Images*. <https://github.com/EdjoLabs/image-match>. Accessed 10 Mar. 2019.

129. *John Resig - Italian Art Computer Vision Analysis*. <https://johnresig.com/research/italian-art-computer-vision-analysis/>. Accessed 10 Mar. 2019.

the Pharos consortium, showed that Pastec.io¹³⁰, an open-source image similarity search tool can provide a very usable set of results for a series of use-cases.

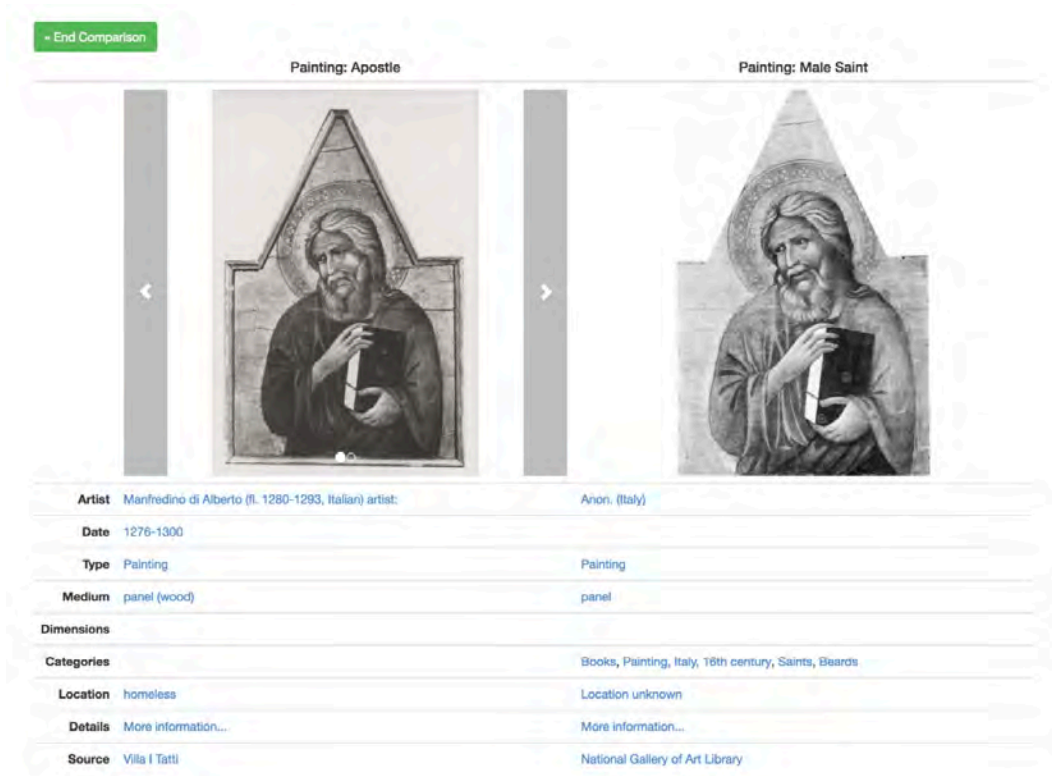


Figure 54: Image matching across institutions in the Pharos test interface

As seen in figure 54, similar images from two institutions are able to be lined up and have their metadata compared. By fine-tuning the results from Pastec, it would be possible to create an artwork disambiguation service that can connect collections of images in museums, photo archives, and institutions with archives that document cultural heritage.

130. Pastec, the Open Source Image Recognition Technology for Your Mobile Apps. <http://pastec.io/>. Accessed 10 Mar. 2019.

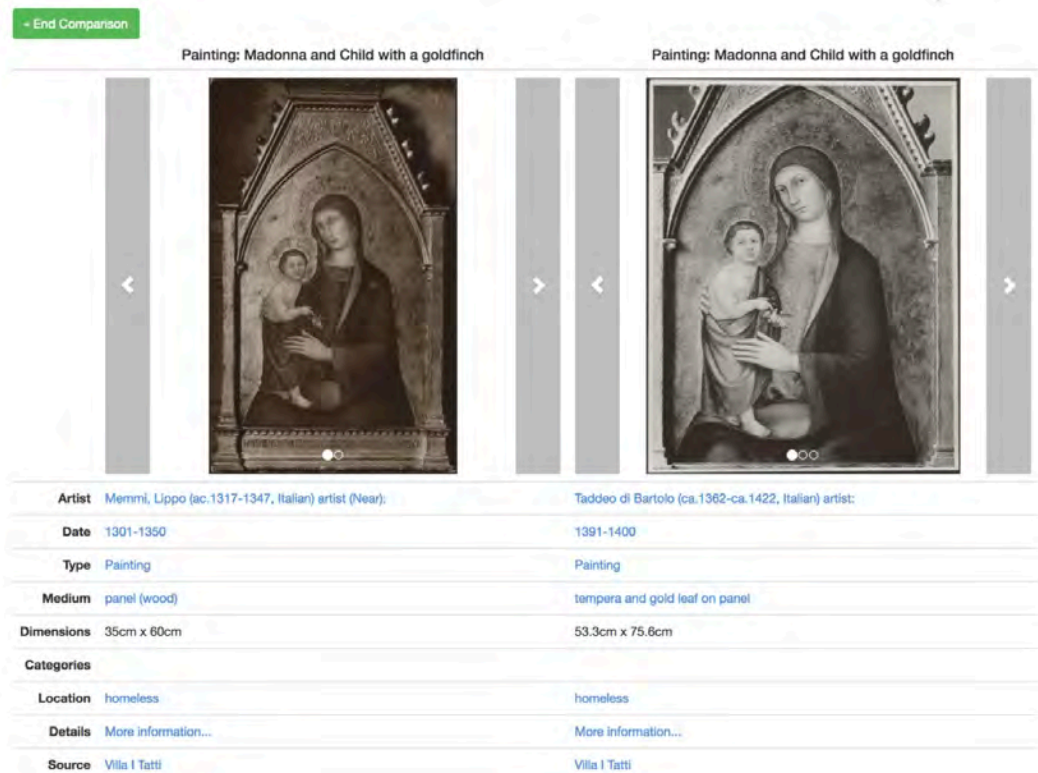


Figure 55: Image matching within one institution in the Pharos test interface

Additionally, the tool can be used internally to programmatically find incongruencies in metadata. As seen in figure 55, a single artwork that was documented in two different photographs was cataloged as having two separate artists and different dates. This is a fairly common cataloging mistake with images of artworks that are lost, where the only data one has to work with is the photograph itself. Using image similarity searching, data about lost artworks can be more easily merged and the histories of these artworks could potentially be reconstructed.



Figure 56: Different artworks being matched

A third use case for searching visually similar images, is scholars investigating copies, or artistic influence. Here the “fuzziness” factor becomes quite wide, and varying perceptions of similarity, based on the tool being used, will have varying results.

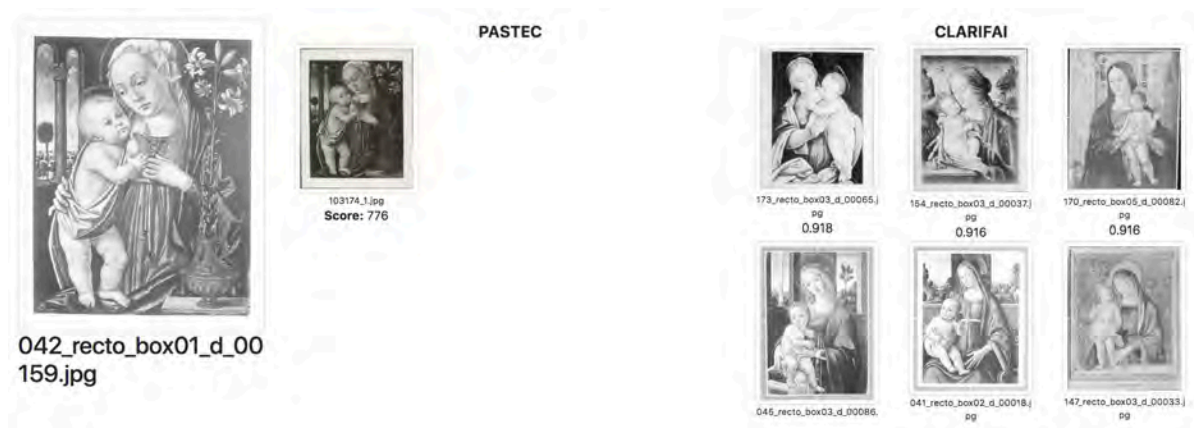


Figure 57: comparing results between Pastec and Clarifai

Given the exact same set of images to search through, figure 57 demonstrates differences between two of these services, Pastec and Clarifai¹³¹. The author conducted an analysis of these services using the FotoIndex and Homeless Paintings collections (which had overlapping images photographed at different times), the results of which were published in a web application that allows for the exploration of these results¹³². While Pastec returns an image of the same artwork represented in a different photograph, Clarifai is not able find that image. In turn, images that are visually similar are returned by the Clarifai service and may nonetheless be useful to art historians. The underlying methodology that Clarifai uses is not disclosed publicly, as it is a commercial service, but it is most likely a Convolutional Neural Network, as it also allows for training based on user input.

Partial Image Match allows for cropped images or details of works to be found within images that contain the whole. There are numerous use cases for this kind of functionality, especially with the prevalence of images that are cropped, or images of artworks that have been split up into pieces, as is often the case with altarpieces and triptychs.

131. Results are viewable here:

Clarifai. <https://clarifai.com/>. Accessed 10 Mar. 2019.

132. Klic, Lukas. *vision.itatti.harvard.edu, an Evaluation of AI Vision Services for the VIT Photo Archive Collection*. <http://vision.itatti.harvard.edu:3000/>. Accessed 10 Feb. 2019.



Figure 58: matching details of artworks to their whole

For this use case, Pastec is once again able to return meaningful results. As seen in figure 58, a small detail of an artwork was matched with relative certainty. Of the services that were tested, Pastec is the only one able to return results for partial image matches that are meaningful. Perhaps this is due to the use of the bag-of-words methodology, which is immune to larger mutations of the image. Extensive test with the Clarifai service showed that it was not at all able to perform on partial image matches.

The results of these tests demonstrate that for the three use-cases described, Pastec is the service that prevails in this front and is a strong contender for integration into a generic Visual Search Semantic Web application. Clarifai, although far less accurate, could also be integrated to provide a more “fuzzy” image matching, although the cost of hosting the images on Clarifai can prove to be prohibitive working with large image sets.

Visual Cataloging

Another use-case that could prove to be disruptive to the field of image cataloging, is to leverage the functionality of partial image match or visually similar image matching to assist in the cataloging of images in batch. With the ability to use images to search for others with the same iconographic theme, such as “Madonna and Child”, this presents a powerful tool for institutions to be able apply metadata to vast numbers of images. This is particularly useful in situations where image metadata is lacking, and could embolden the publishing of historic photograph collections by institutions that have no metadata.



Figure 59: matching iconographical themes

With partial image match functionality, Visual Cataloging can be implemented to search the verso of photographs as well. Annotations, which are often written by the same hand, are often visually similar.

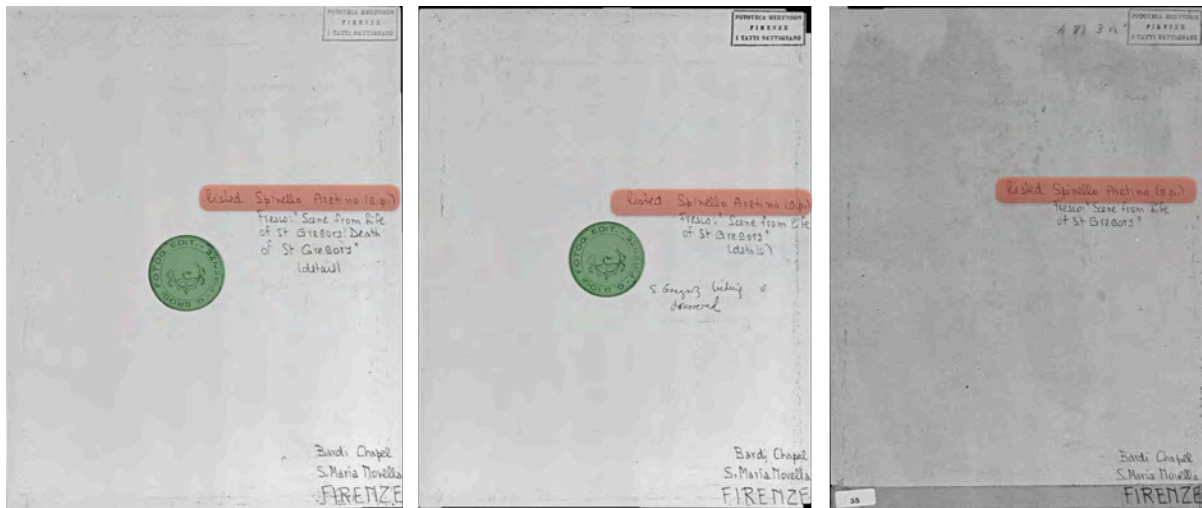


Figure 60: matching stamps and handwritten text on the verso of photographs

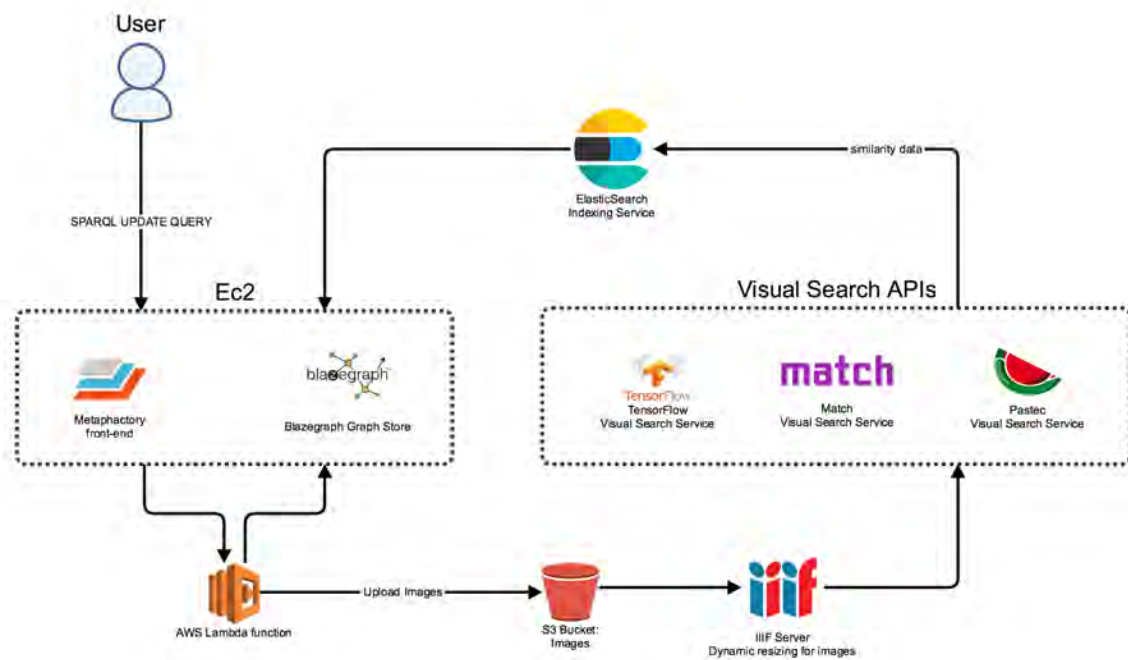
As illustrated by figure 60, this functionality could also be implemented with photographer or institution stamps on the verso of photographs. These kinds of searches could also be very useful for scholars doing research on the history of photography, tracking the movements or works of particular photographers across collections that have not published this metadata as part of their datasets.

System Architecture

In order to accommodate a wide range of visual search functionality with a complex array of results from different sources, this chapter advocates for a semantic framework that allows for the integration of these services. Powered by a graph database with a rich and extensible data

model, the system architecture outlined here will allow for the analysis and interpretation of results from these services in more meaningful ways.

Image similarity services can be integrated into a ResearchSpace instance, which makes the results available to institutions through a SPARQL endpoint, and to non-technical users through a web interface front end. Images can be uploaded in bulk through a SPARQL update query, or through drag-and-drop functionality on the front end. Adding images to be indexed triggers an AWS Lambda¹³³ function that uploads the images from their source to an S3 bucket, resizing the image through a IIF server, and subsequently passing them on to the visual search API's for processing through a SPARQL to REST API component. After parsing the results from the visual search API, they are stored in an ElasticSearch index, and subsequently posted back to Graph database for storage and retrieval once they have been processed.



133. "AWS Lambda – Serverless Compute - Amazon Web Services." *Amazon Web Services, Inc.*, <https://aws.amazon.com/lambda/>. Accessed 30 Jan 2019.

Figure 61: software architecture for Visual Semantic Search

As seen in Figure 61, the AWS Lambda function will handle all of the event management for the process. This is required as there is often some lag time between the moment when an image is uploaded, processed, and the amount of time it will take for the visual search APIs to return a result. This architecture is agnostic to the visual search tool being used, allowing for the system to grow with additional functionality and services over time. When new visual search engines are released, they can be integrated into the architecture without disturbing the other components by simply adding a configuration file that registers the service as a repository, which will handle the translation between SPARQL and the REST API. The Elasticsearch index provides an efficient and speedy lookup of image fingerprints for services such as Match, but is not needed for services such as Pastec and Clarifai that have their own index.

Ontology and Data Model

The underlying ontology and data model is instrumental in the enabling the interoperability of various visual search tools. In order to semantically encode similarity data in the knowledge graph, a simple, but extensible ontology is required to express the various complexities of visual similarity for cultural heritage in a way that is programmatically actionable. Works of cultural heritage pose new challenges and offer opportunities for exploring more abstract concepts of similarity, with models that are flexible enough to accommodate a wide range of use-cases. The data model allows for the description of two images, providing

levels of similarity based on different models or toolsets. While one toolset may use the visual bag-of-words methodology, another could use a CNN to create complex vectors. Each methodology is assigned a single URI, that can have additional properties that describe possible use-cases and how it was implemented. The “score” or weight of similarity can be any numerical value that is specific to that application, and results can easily be sorted numerically through a SPARQL query without a need to define the weight of that score.

To start, a user or institution will need to provide a set of images to parse. Basic metadata will be requested, but requiring the user to provide identifiers for images and artworks (therefore linking images of the same work together).

Field Name	Sample	Type	Importance
Image_URL	http://s3.aws.com/image1234.jpg	Literal	MUST
Image_ID	<http://images.institution.com/ID12345>	URI	MUST
Artwork_ID	<http://artworks.institution.com/ID6789>	URI	MUST
Collection_ID	<http://institution.com/collectionID>	URI	MUST
Work_creation_date_start	1492-01-01	xsd:dateTime	SHOULD
Work_creation_date_end	1492-12-31	xsd:dateTime	SHOULD
Artist_ID	<http://institution.com/artworkID>	URI	COULD
Artist_ID_ULAN	<http://vocab.getty.edu/ulan/500010879>	URI - ULAN	SHOULD

Figure 62: HeritageVision data input

A collection ID, will allow the group of images to be linked together, which will facilitate retrieval at a later date. As optional parameters, users will have the ability to add the ULAN URI of the artist that created the work, or an internal URI, along with a start and end date for the creation, if this data is available. These data will serve at later date for additional analytics and interpretative applications, allowing users to filter results by date and artist. A CSV template that conforms to the data model will be provided, allowing less technically inclined users to input data, together with scripts to convert them to a SPARQL query for ingestion.

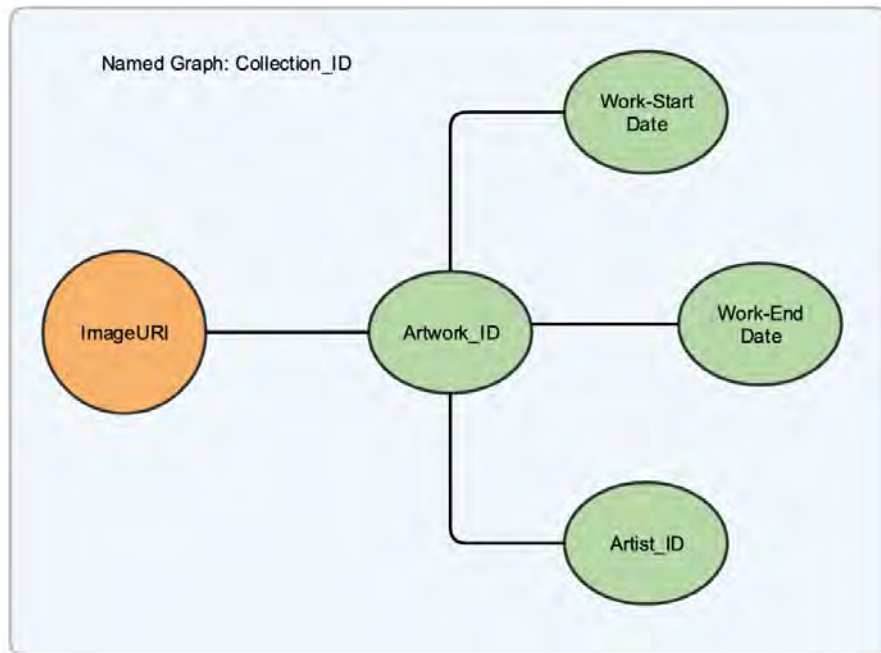


Figure 63: data input model

The data model is kept intentionally simple, in order to lower the complexity of the data input process and encourage users and institutions to upload their images. The uploaded image URI will be passed on to the visual similarity search tools, while the artwork data will be inserted into the graph database, wrapping it in a named graph using the `Collection_ID`.

In order to assess the functional requirements of a data model for visual similarity, some core classes and properties were defined using a fictitious ontology (Visual Similarity, vSim). It was found that at the most basic level, the requirement was to describe a level of similarity between two targets (images), according to a specific methodology. Therefore the data model must have at its core, a node that describes the relationship between two images. In this use case, there is no directionality in terms of similarity, as it is not mutually exclusive to one entity or another and the similarity is bidirectional.

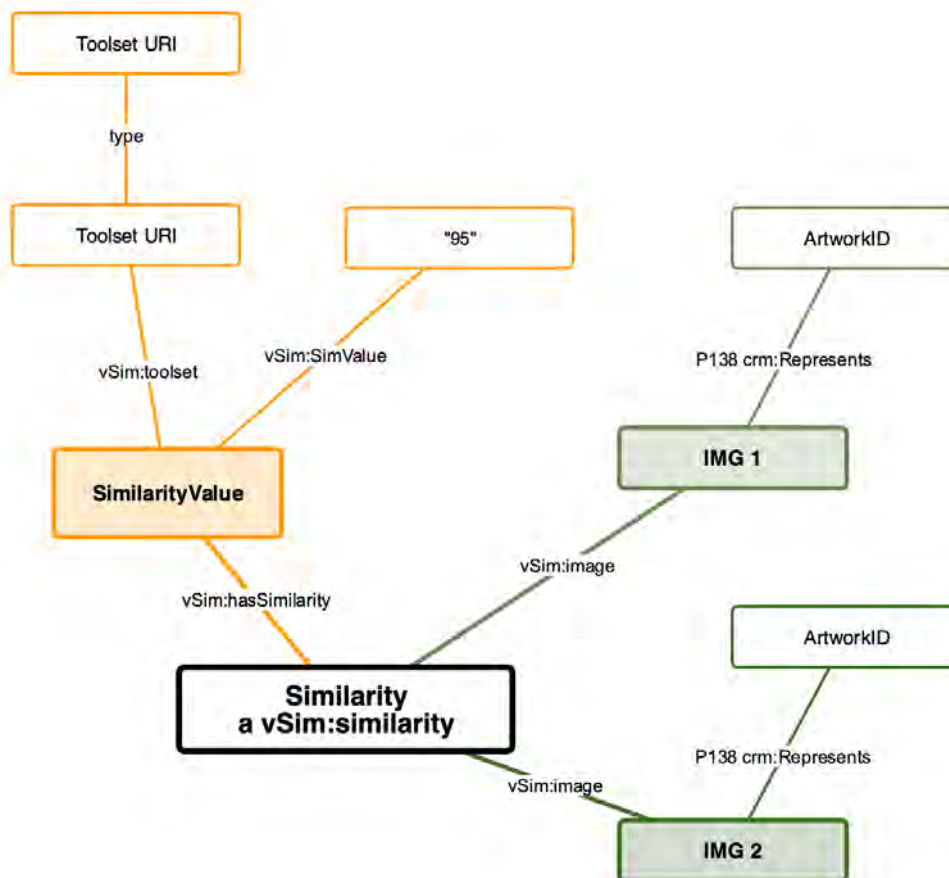


Figure 63: Mock data model to describe visual similarity between two images

A sample of how this data might look in the turtle serialization can be seen below.

```
@prefix vSim: <https://vision.itatti.harvard.edu/resource/ontology/visualsimilarity/> .
```

```
<https://vision.itatti.harvard.edu/resource/similarity/id/s000001>
  a vSim:similarity ;
  vSim:similar_image <https://img1.jpg> ;
  vSim:similar_image <https://img2.jpg> ;
  vSim:similarityValue {URI/UUID} ;
  {URI/UUID} vSim:score "11" .
```


The structure of such a data model would be lightweight enough to provide direct access to the images and a similarity value. With this approach, it would be possible to make multiple statements about the similarity of two images, using more than one service. In order to allow for greater extensibility of the model, an intermediate node describing the tool provides a semantically more rich model for presenting these relationships.

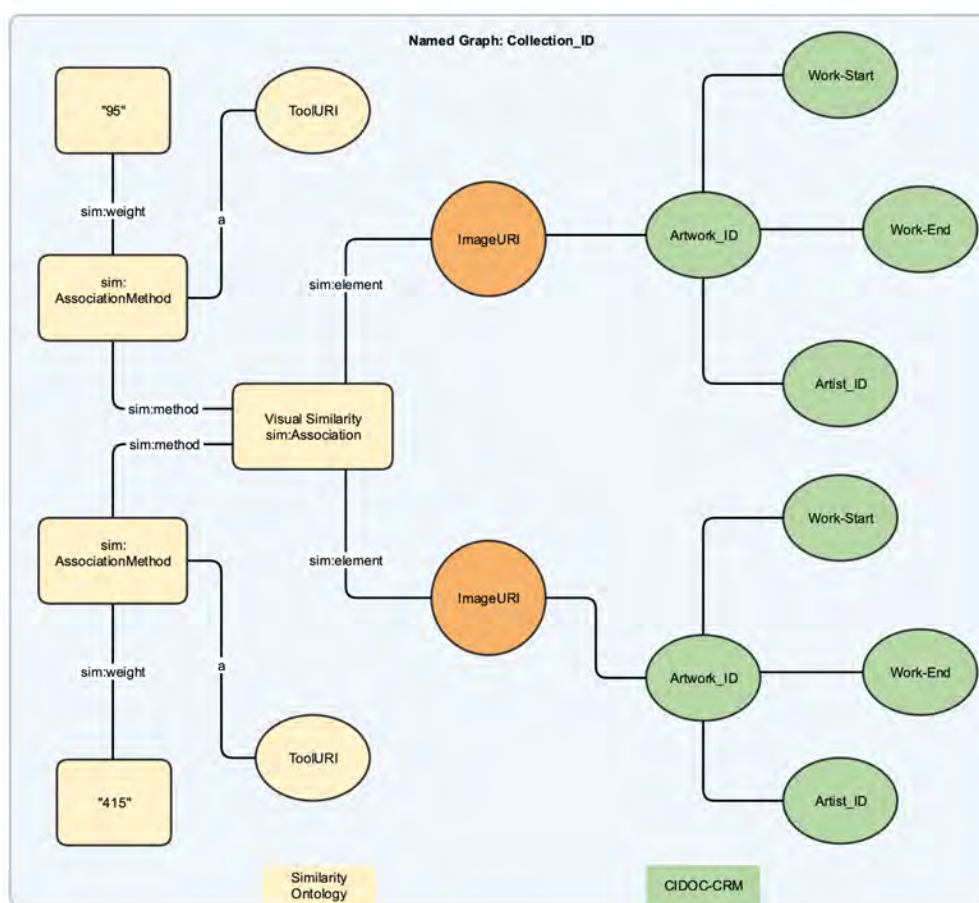


Figure 63: data model for visual similarity

Given the functional requirements of the data model, and to avoid constructing a new ontology, an existing generic ontology was found that covered all of these needs. The *Similarity*

*Ontology*¹³⁴, a lightweight and flexible ontology originally developed for music scores, allows for the creation of similarity statements with the ability to track the provenance of those statements, enabling multiple statements about the similarity of two entities connected to a single similarity node. Although the website documenting the ontology is no longer available (it was last modified in 2010), the internet archive has a copy and the ontology itself is available on Linked Open Vocabularies¹³⁵.

Impact and Future Work

Within the context of the digital collections platform for the Harvard Center, the results of Pastec and Clarifai image similarity searches have been integrated into the datasets of the FotoIndex project, together with those from the Homeless paintings project. Within the artwork template for a given record, a tab displays any visually similar images based on results from these two services, with the ability to navigate to related records. In the future this will be expanded to allow user and cataloger input, enabling them to make assertions themselves regarding the sameness of two artworks. The system architecture described in this chapter, will be implemented into a separate platform that will evolve over several stages, starting with the ability to provide results from Pastec and Clarifai, subsequently expanding to additional services. Functionality to build custom models for visual searches, and the batch application of metadata will be incorporated over time. The platform will run as a standalone service with its own user

134. *The Similarity Ontology*. 16 Jan. 2013, <https://web.archive.org/web/20130116095414/http://kakapo.dcs.qmul.ac.uk/ontology/musim/0.2/musim.html>.

135. *Linked Open Vocabularies (LOV)*. <https://lov.linkeddata.es/dataset/lov/vocabs/sim>. Accessed 10 Mar. 2019.

interface, but since it uses ResearchSpace at its foundation, repository functionality within the software allows any other ResearchSpace instance to seamlessly plug in and fully integrate, simply by registering it as an external SPARQL repository. A catalog of field definitions and template files will allow for any other institution running the platform to quickly add tabs to their artwork records to show images in other repositories that are visually similar.

Over time, the hope is that a growing collection of images will make the platform as an attractive tool for artwork disambiguation. Identifiers for artworks will be used to link records across the web of data, providing the first decentralized repository on the web that can provide such a service. Since the application downloads and stores all images that are input into an S3 bucket, they will be preserved even in the case when host websites change their URI patterns. As more images are added to the collection, the issue of scalability will need to be addressed with some services. Pastec, for example, does not scale well over 1 million images, and therefore alternative services will need to be used, possibly commercial ones (such as TinEye¹³⁶, the original reverse-image search).

Aside from disambiguation services for institutions, the platform could also prove to be transformative for scholars seeking research and collections data from multiple repositories related to a particular artwork. These data, would otherwise be spread out in various silos across the web, with little ability to track them down if not with visual search. While Google image search has been useful to scholars when searching for copies of similar images, most digital collections databases are not indexed by the search giant, as they require the use of a search field to get information back, a technique that Google's web crawlers do not employ. Images and core

136. *TinEye Reverse Image Search*. <https://www.tineye.com/>. Accessed 10 Mar. 2019.

metadata that would allow for some basic filtering would need to be provided by institutions, allowing them to upload their collections for consumption by users and other Linked Data applications via a SPARQL endpoint. The hope is that this platform can serve as an endpoint for artwork disambiguation, assigning an identifier for artworks and act as the linking mechanism between multiple repositories, much as the Getty vocabularies have done for artists (ULAN) and terms to classify various categories of artworks (AAT).

Given the wide range of use-cases, the range of artworks that can benefit from a visual similarity web service, along with the varying results returned from various services, the results will initially serve as a tool for exploring these similarities. Catalogers will be provided with logins to the platform and given the ability to make assertions about the sameness of two images, which will more explicitly connect two artwork identifiers with an owl:sameAs statement. By augmenting machine-generated similarity responses with the assertions of expert users, the dataset over time can serve to build new models that are more finely-tuned to artwork similarity, rather than using generic ones that have been trained on other types of images.

VI - Towards an Open and Collaborative Digital Art History

The discourse around research technology for the humanities is one that is more practical than philosophical. There is general agreement that technology offers a wealth of tools that are of interest to scholars, the principal challenge being in their application and implementation. The field of Digital Humanities has emerged in response to a recognition that humanities research needs to be more interwoven with technology. Although databases of images and library catalogs, together with major search engines have made vast amounts of information more accessible, there has been relatively little evolution in the way that scholars use these systems over the past two decades. These systems generally aggregate data in some form of an index, and a text-based search allows one to search. The result is systems that facilitate what is commonly known in the information science community¹³⁷ as known-item searches. While other advancements such as faceted browsing do allow for the filtering of results that enable one to hone in more closely on results to increase precision, contextual information is still missing from these more traditional information systems. Integration methods offered by the LOD and Semantic Web movement have made great strides to narrowing this gap, but the field is still grappling with the application of these technologies. Due to the relative slow pace in

137. Lee, Jin Ha, et al. "Known-Item Search: Variations on a Concept." *Proceedings of the American Society for Information Science and Technology*, vol. 43, no. 1, 2006, pp. 1–17. *Wiley Online Library*, doi:[10.1002/meet.14504301126](https://doi.org/10.1002/meet.14504301126).

advancement of information systems, much of the community of humanities scholars that use these systems have become somewhat complacent in their research practices, still resorting to major search engines as their primary first step in the research process. While the efficiency of search engines has grown over time and proven to be transformative in comparison to the research practices of the 20th century and even twenty years ago when Yahoo.com was still employing people to manually catalog the web, scholars are still using these systems primarily for basic access. A transition to collaborative research environments where personal archives of data and research are woven together with data across the web, will require a substantial shift in mindset and methodology. In the humanities and sciences alike, scholars have a tendency to amass large archives of data, PDF's, and images on their personal computers to support their research practices. The concept of sharing these data without offering a mechanism for citing them, is generally not conducive to academia. In order to support and encourage a more collaborative and open Digital Art History, institutions should begin to recognize the publishing efforts of scholars beyond the traditional means of a printed article or book. The concept of micropublishing has the potential to alleviate some of these issues, and platforms such as ScienceMatters¹³⁸ have begun to support the publishing of small scientific observations that may not warrant a full article. These streamlined publishing processes facilitate the adoption of such systems, but they still lack the ability to publish structured, interactive and programmatically actionable data that can be interwoven into these scholarly articles. The ResearchSpace system serves to address this gap, where the full lifecycle of research process can be managed in a single environment, linking scholarly articles with research data, with the ability to collaborate with

138. *ScienceMatters* | *ScienceMatters*. <https://www.sciencematters.io/>. Accessed 17 Mar. 2019.

others. In the Digital Humanities where the research questions are often too large for a single individual, this system has the potential to be transformative to field.

Impact

The field has recently seen a surge in calls for publishing research and collections data that allow for reusability and integration¹³⁹, as there is growing awareness of a need to be able to link resources across the web of data. As outlined in the Ruben Verborgh's article *The Semantic Web identity crisis: in search of the trivialities that never were*¹⁴⁰, there is a discrepancy between the theoretical problems that the semantic web addresses and their real-world implementations. The assumption is that since many problems have been addressed at a theoretical level, the rest is a software engineering problem. Even within the Semantic Web community, there is agreement that Tim Berners-Lee's vision of the Semantic Web as outlined in his Scientific American article in 2001 has not been realized¹⁴¹, and that this is the result of both the complexity of publishing semantically enriched data, a lack of agreement at the implementation level, as well as a lack of engineers building software to provide real-world implementations. The methodologies shared in this project aim to assist institutions and individuals in lowering the barriers for achieving this task. Since the ResearchSpace project is still in its early stages, more work needs to be done on sharing the know-how of how to publish data in such ecosystems, and is one of the principal

139. *Linked Research*. <https://linkedresearch.org/>. Accessed 21 May 2019.

140. Verborgh, Ruben, and Miel Vander Sande. "The Semantic Web Identity Crisis: In Search of the Trivialities That Never Were." *Semantic Web*, no. pre-press, <http://www.semantic-web-journal.net/content/semantic-web-identity-crisis-search-trivialities-never-were>. Accessed 21 May 2019.

141. Hogan, Aidan. "The Semantic Web: Two Decades On." *The Semantic Web*, pre-press, p. 14.

outcomes of the FotoIndex project. The methodologies presented here can serve as a guide to other individuals and institutions seeking to publish similar collections data as part of larger Digital Humanities projects that wish to leverage Linked Data and Semantic Web technologies to interlink and enrich collections, both by interlinking identifiers and by leveraging Computer Vision services in a Semantic Web environment. By providing a high-level analysis of the transformation process, sharing methodologies and tools for the cleaning and transformation of legacy data to RDF, it serves to act as a stepping stone away from silos of research and collections data, towards an open, interlinked, and collaborative future that enables researchers to interact with one another before, during and after the research process. Additionally, the dataset generated from this project can serve as the foundation for further research, and can be leveraged to enrich other collections on web, in particular those that include data related to the provenance of Renaissance artworks.

The methodologies of the FotoIndex project have already made a substantial impact on the field, as the project has been presented at the Art Libraries Society of North America¹⁴², the Getty Research Institute in California¹⁴³, The Biblioteca Hertziana¹⁴⁴, the International Conference of Art Libraries¹⁴⁵, among others. It has served as both a proof of concept and as inspiration for other institutions seeking to publish archival material, without the financial means to do so in traditional ways with catalogers. The Getty Research Institute (GRI) recently

142. Klic, Lukas. *(Mass)Digitizing the Berenson Photo Archive at Villa I Tatti: Metadata Creation, Enrichment, and Discovery*. ARLIS/NA (Art Libraries Society of North America), New York City, USA.

143. Klic, Lukas. *(Mass)Digitizing the Berenson Photo Archive: From Metadata Creation & Enrichment to Discovery*. Getty Research Institute. Los Angeles, California.

144. Klic, Lukas. *Digital Scholarly Publishing: Moving Beyond the Printed Book*. Biblioteca Hertziana, Max Planck Institute for Art History. Rome, Italy.

145. Klic, Lukas. *Integrating Digital Collections. PHAROS: The International Consortium of Photo Archives*. 8th international Conference of Art Libraries, Amsterdam, Netherlands.

announced their plan to digitize nearly one million photographs without cataloging, taking inspiration from the methodologies presented by the author. The GRI project similarly involves only digitization and the creation of a minimal index for these records, leaving the rest up to Computer Vision, Machine Learning, and other technologies that will automate the metadata generation process. In a related initiative, negotiations are currently underway with the Andrew W. Mellon foundation to fund a project from the PHAROS consortium¹⁴⁶, where the author serves as the Technical Architect, also based on methodologies from the FotoIndex project. This project involves the transformation of existing datasets from five institutions across Europe and North America, linking the datasets to common vocabularies (Getty ULAN, ATT, WikiData, VIAF, GeoNames, etc.) using the CIDOC CRM ontology, all within the ResearchSpace platform. Here, the Matchmaker application will be used to align those archival data with internal and external datasets, in order to allow for the disambiguation of entities and to enrich them further with contextual data. Under this project, the visual search for the Semantic Web architecture will be tested, and will serve as a linking tool to connect images of the same artworks across repositories.

While in recent years there has been a lull in digitization efforts and funding to support such initiatives, the FotoIndex project serves as a revitalization mechanism that has inspired additional institutions to follow. The risk being, that if many of these archives are not digitized, the institutional memory will be soon lost and these collections may suffer a fate of never being accessed again. Finally, expanding the digital collections of Villa I Tatti is of exceptional benefit to the field of early modern studies and the humanities across the board, giving scholars globally

146. PHAROS: *The International Consortium of Photo Archives*. <http://pharosartresearch.org/>. Accessed 17 Mar. 2019.

free and open access to over a century of the highest quality scholarship in the field, generated at the Harvard Center.

Future Work

The FotoIndex project captured and published the most important historical photographs in the Villa I Tatti collection up until the 1980's. Since the photo archive is actively growing, much material (nearly 100,000 new photos) have been added since. New systems have already been put in place to be able to digitize these materials in the most efficient way possible, using economical form-feeding scanners recently released by Epson, together with a large overhead scanner where multiple larger and more fragile photos can be placed on the surface, and are automatically cropped into separate images. This new digitization workflow will be powered by another MatchMaker module that will ensure the process is as seamless as possible. Scanning operators will process one box of photos at a time, first inputting metadata from the box in an auto-suggest form (artist and provenance information). Feeding a photo into the scanner will automatically create an identifier for the object and print out a barcode, which can then be applied to the physical object. Since the scanning process will be done in a controlled environment, the images will be immediately cropped down and the operator will have the chance to verify its accuracy on the fly. Thanks to the existing dataset of artists and institutions,

the auto-lookup functionality will ensure that subsequent reconciliation processes are not necessary. Additionally, since some of the images in the FotoIndex were of poor quality (they intended only to be reference images), Pastec will be used to auto-match new scans to older ones, and link these digital surrogates to a single photograph record, as well as a record for a particular artwork. Once the new digitization workflow is complete, the Matchmaker application will be refined to be collection-agnostic and published on GitHub, allowing any institution to use the same (or similar) scanner to digitize photographs, load up their collections data, link images to metadata, reconcile the data, and then export it for subsequent transformation to RDF using the CIDOC-CRM ontology with Memory Mapping Manager and the X3ML engine.

Platform-wise, three different systems will serve different functionalities, all built on top of ResearchSpace. The collections portal for the Harvard Center (collection.itatti.harvard.edu) will host the archival collections of I Tatti, including the images from photo archive, the archive, and digitized library materials. This platform will also be used to catalog new items, and enrich the metadata of existing collections, thus allowing for a transition away from the current SharedShelf system. A separate visual search platform (LinkedOpenImages.com) as outlined in chapter five will initially only provide identifiers for artworks on the web by matching up artworks, making a statement about their similarity. Visual cataloging and custom model building functionality will be introduced as well, allowing institutions that lack metadata to enrich their collections through visual search and the batch application of metadata.

Finally, ArtResearch.net will serve as a research platform that will allow scholars browse the collections from multiple institutions through federated searching, and augment this data with their own research. It will also allow them to build their own personal digital libraries, being able

to upload images of archival documents, transcribe them, and enable semantically enriched annotations. This same functionality will also allow the platform to serve as a digital publishing platform, where scholars can write a narrative about a given topic in the form of an article or micropublication, annotating entities and their related network of relationships. ArtResearch.net will focus on building new knowledge, while allowing users to tap into the resources of other Linked Data environments across the web.

Although most of the implementation work on these systems has been completed, the final publishing of these platforms is currently being rewritten for the release of ResearchSpace 3.0, which was released in May of 2019. This release has substantial architectural changes that will allow for the long-term management of the platform to be streamlined, as the storage mechanisms for data, templates, and field definitions has been decoupled from platform itself, allowing new developments (such as the semantic annotation of text documents¹⁴⁷) to be integrated without needing to rebuild all of the configurations and templates.

Conclusion

The process of building out digital collections and supporting research infrastructure is topic of great interest to scholars and cultural heritage institutions. By creating working environments that are constantly growing with new research being contributed regularly, the longevity of digital projects can be ensured for years to come. Many digital projects from the early 2000's have now suffered a very unpleasant fate: system architectures and data structures

147. *Semantic Digital Publishing - Semantic Digital Publishing - Consortium for Open Research Data in the Humanities*. <https://wiki.cordh.net/display/SDP/Semantic+Digital+Publishing>. Accessed 17 Mar. 2019.

that were too rigid and non portable, combined with a lack of institutional will, have resulted in many projects that cease to exist, together with their related research data. Many of these earlier projects have managed to stay online only by converting them to static HTML websites, where any functionality (including basic search) has been stripped away. While keeping software up-to-date and patched is an important step, even more important is designing the data models in a way that allows them to be portable to other systems. With the advent of RDF, data portability is greatly facilitated, as the data itself, the model, and ontology are all clearly navigable, and stored in one location. If the architecture of the data model is properly encoded, all meaning that is necessary to be able to read and interpret it, is contained within this structure, a feature that relational databases cannot boast. In this way, the collection becomes resilient to issues of long-term preservation. This focus on data architecture is key, as most computer scientists agree, in a period of ten years (or less) most systems become obsolete. Building data structures that are system agnostic, will allow them to live on, either in the form of raw data or with a different system architecture and user interface.

This project strongly advocates for institutions to adopt a single system for the publishing of research and collections data, both to facilitate maintenance but also integration. Research centers and universities that have multiple digital projects struggle to keep software patched and up-to-date, as maintaining legacy digital projects over time becomes increasingly cumbersome. These projects also remain frozen in time if they do not offer the functionality to allow for a community of scholars to augment that data. As personnel move around to other institutions, the know-how for how to maintain these projects is lost, resulting in digital publishing environments that are ephemeral. Integrating multiple projects under a single system architecture that has a

broad set of use-cases and allows for custom interfaces to be built around datasets, addresses this issue for the cultural heritage community. Equally important is the need to build communities of individuals and institutions around these systems, to ensure their adoption and growth. For this reason, the author is a founding partner in the Consortium for Open Research Data in the Humanities (cordh.net¹⁴⁸), and has been instrumental in galvanizing a community of institutions committed to expanding and supporting the ResearchSpace system, using the CIDOC-CRM as the underlying data model. Although the original ResearchSpace system was built to support the collections of the British Museum, the author has worked closely with the developers of the software and the British Museum to make it more collection agnostic. The 3.0 release provides a more generic user interface that allows RDF data to be loaded together with filed definitions that define the relationships between entities, allowing for the instant visualization of research and collections data. Annotation functionality allows scholars to enrich collections data with observations or annotations, creating a digital environment that is open, inclusive, and collaborative, providing a community-driven platform that offers new insights into the history of culture.

148. "Consortium for Open Research Data in the Humanities." *Consortium for Open Research Data in the Humanities*, <https://www.cordh.net/>. Accessed 17 Mar. 2019.

Bibliography

3M. <http://www.ics.forth.gr/isl/3M/>. Accessed 9 Mar. 2019.

Abbott, Alison. "The 'Time Machine' Reconstructing Ancient Venice's Social Networks." *Nature News*, vol. 546, no. 7658, June 2017, p. 341. *www.nature.com*, doi:[10.1038/546341a](https://doi.org/10.1038/546341a).

About the American Art Collaborative | American Art Collaborative. <http://americanartcollaborative.org/about/>. Accessed 22 Feb. 2019.

About the Research Institute (Getty Research Institute). <http://www.getty.edu/research/institute/>. Accessed 23 Feb. 2019.

Acheson, Elise, et al. "A Quantitative Analysis of Global Gazetteers: Patterns of Coverage for Common Feature Types." *Computers, Environment and Urban Systems*, vol. 64, July 2017, pp. 309–20. *ScienceDirect*, doi:[10.1016/j.compenvurbsys.2017.03.007](https://doi.org/10.1016/j.compenvurbsys.2017.03.007).

Amazon.Com: What Is an Image? (The Stone Art Theory Institutes) (9780271050645): James Elkins, Maja Naef: Books. <https://www.amazon.com/What-Image-Stone-Theory-Institutes/dp/0271050640>. Accessed 10 Mar. 2019.

American Art Collaborative. <http://americanartcollaborative.org/>. Accessed 23 Feb. 2019.

Apache Solr -. <https://lucene.apache.org/solr/>. Accessed 8 Mar. 2019.

ArcGIS Online | Interactive Maps Connecting People, Locations & Data. <https://www.arcgis.com/index.html>. Accessed 23 Feb. 2019.

ArtFrame - LD4P Public Website - DuraSpace Wiki. <https://wiki.duraspace.org/display/LD4P/ArtFrame>. Accessed 23 Feb. 2019.

ArtPi - Artrendex. <http://www.artrendex.com/artpi>. Accessed 10 Mar. 2019.

“AWS Lambda – Serverless Compute - Amazon Web Services.” *Amazon Web Services, Inc.*, <https://aws.amazon.com/lambda/>. Accessed 30 May 2019.

Baseline Patterns. <https://linked.art/model/base/>. Accessed 23 Feb. 2019.

Berardi, Elena. *NORMATIVA F - FOTOGRAFIA: STRUTTURAZIONE DEI DATI E NORME DI COMPILAZIONE. 4.0*, ISTITUTO CENTRALE PER IL CATALOGO E LA DOCUMENTAZIONE, 2015, p. 180. Zotero, <http://www.iccd.beniculturali.it/getFile.php?id=4479>.

Berenson, Bernard. *I disegni dei pittori fiorentini*. Electa editrice, 1961.

Berners-Lee, Tim, et al. “The Semantic Web: A New Form of Web Content That Is Meaningful to Computers Will Unleash a Revolution of New Possibilities.” *ScientificAmerican.Com*, May 2001.

Billey, Amber M., et al. *The Outcome of the ArtFrame Project, a Domain-Specific BIBFRAME Exploration*. 2018. *academiccommons.columbia.edu*, doi:[10.7916/D8281M24](https://doi.org/10.7916/D8281M24).

Bing Visual Search Developer Platform. <https://www.bingvisualsearch.com/docs>. Accessed 10 Mar. 2019.

Boer, Victor de, et al. “Supporting Linked Data Production for Cultural Heritage Institutes: The Amsterdam Museum Case Study.” *The Semantic Web:*

Research and Applications, Springer, Berlin, Heidelberg, 2012, pp. 733–47.
link.springer.com, doi:[10.1007/978-3-642-30284-8_56](https://doi.org/10.1007/978-3-642-30284-8_56).

“British Museum Publishes Its Collection Semantically.” *British Museum*,
https://www.britishmuseum.org/about_us/news_and_press/press_releases/2011/semantic_web_endpoint.aspx. Accessed 22 Feb. 2019.

Callimachus - Data-Driven Applications Made Easy. <http://callimachusproject.org/>. Accessed 10 Mar. 2019.

Cantaloupe Image Server :: Home. <https://medusa-project.github.io/cantaloupe/>. Accessed 2 Mar. 2019.

Carbon LDP. <https://carbonldp.com/>. Accessed 10 Mar. 2019.

Carlson, Scott, and Amber Seely. “Using OpenRefine’s Reconciliation to Validate Local Authority Headings.” *Cataloging & Classification Quarterly*, vol. 55, no. 1, Jan. 2017, pp. 1–11. *Taylor and Francis+NEJM*, doi:
[10.1080/01639374.2016.1245693](https://doi.org/10.1080/01639374.2016.1245693).

Chiu, Jeff. *OpenRefine Reconciliation Services for VIAF, ORCID, and Open Library + Framework for Creating More.: Codeforkjeff/Conciliator*. 2016. 2019. *GitHub*,
<https://github.com/codeforkjeff/conciliator>.

Ciotti, Fabio, and Francesca Tomasi. “Formal Ontologies, Linked Data, and TEI Semantics.” *Journal of the Text Encoding Initiative*, no. Issue 9, Sept. 2016.
journals.openedition.org, doi:[10.4000/jtei.1480](https://doi.org/10.4000/jtei.1480).

Clarifai. <https://clarifai.com/>. Accessed 10 Mar. 2019.

Cloudy Vision, a Comparison of Image Tagging from Various Vendors for the VIT Photo Archive Collection. https://lklic.github.io/compare_vision/output/output.html. Accessed 10 Mar. 2019.

Compatible Models & Collaborations | CIDOC CRM. <http://www.cidoc-crm.org/collaborations>. Accessed 9 Mar. 2019.

“Consortium for Open Research Data in the Humanities.” *Consortium for Open Research Data in the Humanities*, <https://www.cordh.net/>. Accessed 23 Feb. 2019.

Creating URIs | Data.Gov.Uk. 15 July 2017, <https://web.archive.org/web/20170715074122/https://data.gov.uk/resources/uris>.

CRIA - Committee to Rescue Italian Art. <https://cria.itatti.harvard.edu/>. Accessed 24 Feb. 2019.

Crowley, Elliot J., and Andrew Zisserman. “In Search of Art.” *Computer Vision - ECCV 2014 Workshops*, edited by Lourdes Agapito et al., Springer International Publishing, 2015, pp. 54–70.

Crystal_ball: Scalable Reverse Image Search Built on Kubernetes and Elasticsearch: Dsys/Match. 2016. Distributed Systems, 2019. *GitHub*, <https://github.com/dsys/match>.

Cultural Objects Name Authority (Getty Research Institute). <http://www.getty.edu/research/tools/vocabularies/cona/>. Accessed 10 Mar. 2019.

Daquino, Marilena, et al. *Zeri Photo Archive RDF Dataset.* Alma Mater Studiorum - Università di Bologna, 2016. *DataCite*, doi:[10.6092/unibo/amsacta/5157](https://doi.org/10.6092/unibo/amsacta/5157).

de Boer, Victor, et al. "Supporting Linked Data Production for Cultural Heritage Institutes: The Amsterdam Museum Case Study." *The Semantic Web: Research and Applications*, edited by Elena Simperl et al., Springer Berlin Heidelberg, 2012, pp. 733–47.

Deep Video Analytics. <https://www.deepvideoanalytics.com/>. Accessed 10 Mar. 2019.

Diamond. <https://diamond.timemachine.eu/>. Accessed 10 Mar. 2019.

Digilib - The Digital Image Library –. <http://digilib.sourceforge.net/index.html>. Accessed 2 Mar. 2019.

Dijkshoorn, Chris, et al. "The Rijksmuseum Collection as Linked Data." *Semantic Web*, vol. 9, no. 2, Jan. 2018, pp. 221–30. *content-iospress-com.ezpprod1.hul.harvard.edu*, doi:[10.3233/SW-170257](https://doi.org/10.3233/SW-170257).

Dodds, Leigh. "Managing RDF Using Named Graphs." *Lost Boy*, 5 Nov. 2009, <https://blog.ldodds.com/2009/11/05/managing-rdf-using-named-graphs/>.

Doerr, Martin. "Ontologies for Cultural Heritage." *Handbook on Ontologies*, edited by Steffen Staab and Rudi Studer, Springer Berlin Heidelberg, 2009, pp. 463–86. *Springer Link*, doi:[10.1007/978-3-540-92673-3_21](https://doi.org/10.1007/978-3-540-92673-3_21).

Drucker, Johanna. "Is There a 'Digital' Art History?" *Visual Resources*, vol. 29, no. 1–2, June 2013, pp. 5–13. *Taylor and Francis+NEJM*, doi:[10.1080/01973762.2013.761106](https://doi.org/10.1080/01973762.2013.761106).

EdjoLabs/Image-Match: Quickly Search over Billions of Images. <https://github.com/EdjoLabs/image-match>. Accessed 10 Mar. 2019.

Elgammal, Ahmed, and Babak Saleh. "Quantifying Creativity in Art Networks." *ArXiv:1506.00711 [Cs]*, June 2015. *arXiv.org*, <http://arxiv.org/abs/1506.00711>.

Elkins, James, and Maja Naef. *What Is an Image?* 2011.

Enríquez, J. G., et al. "Entity Reconciliation in Big Data Sources: A Systematic Mapping Study." *Expert Systems with Applications*, vol. 80, Sept. 2017, pp. 14–27. *ScienceDirect*, doi:[10.1016/j.eswa.2017.03.010](https://doi.org/10.1016/j.eswa.2017.03.010).

Federal Agencies Digital Guidelines Initiative. <http://www.digitizationguidelines.gov/>. Accessed 2 Mar. 2019.

Fuzzy Times on Space-Time Volumes - IEEE Conference Publication. <https://ieeexplore-ieee-org.ezp-prod1.hul.harvard.edu/document/7058180>. Accessed 3 Oct. 2018.

Gaehtgens, Thomas W. "Thoughts on the Digital Future of the Humanities and Art History." *Visual Resources*, vol. 29, no. 1–2, June 2013, pp. 22–25. *Crossref*, doi:[10.1080/01973762.2013.761110](https://doi.org/10.1080/01973762.2013.761110).

Gemma Boon. <http://network.icom.museum/cidoc/blog/gemma-boon/>. Accessed 22 June 2017.

Gephi - The Open Graph Viz Platform. <https://gephi.org/>. Accessed 23 Feb. 2019.

Getty Vocabularies. <http://vocab.getty.edu/>. Accessed 22 Feb. 2019.

Gonano, Ciro Mattia, et al. "Zeri e LOD: Extracting the Zeri Photo Archive to Linked Open Data: Formalizing the Conceptual Model." *Proceedings of the 14th*

ACM/IEEE-CS Joint Conference on Digital Libraries, IEEE Press, 2014, pp. 289–298. ACM Digital Library, <http://dl.acm.org/citation.cfm?id=2740769.2740820>.

“GraphDB Free Download.” *Ontotext*, <https://www.ontotext.com/free-graphdb-download/>. Accessed 4 Mar. 2019.

Halpin, Harry, et al. *When Owl:SameAs Isn't the Same: An Analysis of Identity Links on the Semantic Web*. p. 4.

Harlow, Christina. *GeoNames Reconciliation Service for OpenRefine/LODRefine/Google Refine: Cmharlow/Geonames-Reconcile*. 2015. 2019. *GitHub*, <https://github.com/cmharlow/geonames-reconcile>.

Haslhofer, Bernhard, and Antoine Isaac. “Data.Europeana.Eu: The Europeana Linked Open Data Pilot.” *International Conference on Dublin Core and Metadata Applications*, vol. 0, Sept. 2011, pp. 94–104.

Heath, Tom, and Christian Bizer. “Linked Data: Evolving the Web into a Global Data Space.” *Synthesis Lectures on the Semantic Web: Theory and Technology*, vol. 1, no. 1, Feb. 2011, pp. 1–136. *CrossRef*, doi:[10.2200/S00334ED1V01Y201102WBE001](https://doi.org/10.2200/S00334ED1V01Y201102WBE001).

Hernandez, David. “Visual Search by Machine Box.” *Machine Box*, 7 Oct. 2017, <https://blog.machinebox.io/visual-search-by-machine-box-eb30062d8abe>.

“Hidden Collections • CLIR.” *CLIR*, <https://www.clir.org/hiddencollections/>. Accessed 6 Feb. 2019.

Hogan, Aidan. “The Semantic Web: Two Decades On.” *The Semantic Web*, no. pre-press, p. 14.

Homeless Paintings of the Italian Renaissance | I Tatti | The Harvard University Center for Italian Renaissance Studies. <http://itatti.harvard.edu/berenson-library/collections/photograph-archives/homeless-paintings>. Accessed 24 Feb. 2019.

How Art History Is Failing at the Internet | The Daily Dot. <https://www.dailydot.com/via/art-history-failing-internet/>. Accessed 3 Oct. 2018.

Hyvönen, Eero, et al. “Linked Data Finland: A 7-Star Model and Platform for Publishing and Re-Using Linked Datasets.” *The Semantic Web: ESWC 2014 Satellite Events*, edited by Valentina Presutti et al., Springer International Publishing, 2014, pp. 226–30.

ICOM. <https://icom.museum/en/>. Accessed 19 Mar. 2019.

ICS - CRMinf: The Argumentation Model. https://www.ics.forth.gr/isl/index_main.php?l=e&c=713. Accessed 9 Mar. 2019.

Identifying Similar Images with TensorFlow. <http://douglasduhaime.com/posts/identifying-similar-images-with-tensorflow.html>. Accessed 9 July 2018.

IEEE Xplore Full Text PDF. <http://ieeexplore.ieee.org/ielx7/7527113/7532277/07533051.pdf?tp=&arnumber=7533051&isnumber=7532277>. Accessed 10 Mar. 2019.

Image API 2.1.1 — IIIF | International Image Interoperability Framework. <https://iiif.io/api/image/2.1/#region>. Accessed 2 Mar. 2019.

Impett, Leonardo, and Sabine Süssstrunk. “Pose and Pathosformel in Aby Warburg’s Bilderatlas.” *Computer Vision – ECCV 2016 Workshops*, edited by Gang Hua and Hervé Jégou, Springer International Publishing, 2016, pp. 888–902.

Introduction to Solr Indexing | *Apache Solr Reference Guide 6.6*. https://lucene.apache.org/solr/guide/6_6/introduction-to-solr-indexing.html. Accessed 8 Mar. 2019.

John Resig - Building an Art History Database Using Computer Vision. <https://johnresig.com/blog/building-art-history-database-computer-vision/>. Accessed 10 Mar. 2019.

John Resig - Italian Art Computer Vision Analysis. <https://johnresig.com/research/italian-art-computer-vision-analysis/>. Accessed 10 Mar. 2019.

JSDoc: Home. <http://linkeddata.github.io/rdfliib.js/doc/>. Accessed 10 Mar. 2019.

JSON-LD 1.1. <https://www.w3.org/2018/jsonld-cg-reports/json-ld/>. Accessed 23 Feb. 2019.

JSON-LD and Why I Hate the Semantic Web | *The Beautiful, Tormented Machine*. <http://manu.sporny.org/2014/json-ld-origins-2/>. Accessed 23 Feb. 2019.

Karayev, Sergey, et al. "Recognizing Image Style." *ArXiv:1311.3715 [Cs]*, Nov. 2013. *arXiv.org*, <http://arxiv.org/abs/1311.3715>.

Karma: A Data Integration Tool. <http://usc-isi-i2.github.io/karma/>. Accessed 9 Mar. 2019.

Kienle, Miriam. "Between Nodes and Edges: Possibilities and Limits of Network Analysis in Art History." *Artl@s Bulletin*, vol. 6, no. 3, Nov. 2017, <http://docs.lib.purdue.edu/artlas/vol6/iss3/1>.

Klic, Lukas. *Digital Scholarly Publishing: Moving Beyond the Printed Book*. Biblioteca Hertziana, Max Planck Institute for Art History. Rome, Italy.

Klic, Lukas, Jonathan K. Nelson, et al. "Florentine Renaissance Drawings: A Linked Catalog for the Semantic Web." *Art Documentation: Journal of the Art Libraries Society of North America*, vol. 37, no. 1, Mar. 2018, pp. 33–43. www-journals-uchicago-edu.ezp-prod1.hul.harvard.edu (Atypon), doi:[10.1086/697276](https://doi.org/10.1086/697276).

Klic, Lukas. *Integrating Digital Collections. PHAROS: The International Consortium of Photo Archives*. 8th international Conference of Art Libraries, Amsterdam, Netherlands.

---. (Mass)Digitizing the Berenson Photo Archive at Villa I Tatti: Metadata Creation, Enrichment, and Discovery. ARLIS/NA (Art Libraries Society of North America), New York City, USA.

---. (Mass)Digitizing the Berenson Photo Archive: From Metadata Creation & Enrichment to Discovery. Getty Research Institute. Los Angeles, California.

Klic, Lukas, Matt Miller, et al. *The Code4Lib Journal – The Drawings of the Florentine Painters: From Print Catalog to Linked Open Data*. no. 38, Oct. 2017, <https://journal.code4lib.org/articles/12902>.

Klic, Lukas. *vision.itatti.harvard.edu, an Evaluation of AI Vision Services for the VIT Photo Archive Collection*. <http://vision.itatti.harvard.edu:3000/>. Accessed 10 Mar. 2019.

Known-Item Search: Variations on a Concept - Lee - 2006 - Proceedings of the American Society for Information Science and Technology - Wiley Online Library. <https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/full/10.1002/meet.14504301126>. Accessed 11 Mar. 2019.

Kohle, Hubertus. "Kunstgeschichte und Digital Humanities. Einladung zu einer Debatte." *Zeitschrift für Kunstgeschichte*, July 2016, pp. 151–54.

Lee, Jin Ha, et al. "Known-Item Search: Variations on a Concept." *Proceedings of the American Society for Information Science and Technology*, vol. 43, no. 1, 2006, pp. 1–17. *Wiley Online Library*, doi:[10.1002/meet.14504301126](https://doi.org/10.1002/meet.14504301126).

Lian, Xiang, et al. "K-Nearest Keyword Search in RDF Graphs." *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 22, Oct. 2013, pp. 40–56. *Crossref*, doi:[10.1016/j.websem.2013.08.001](https://doi.org/10.1016/j.websem.2013.08.001).

Lincoln, Matthew. "Continuity and Disruption in European Networks of Print Production, 1550-1750." *Artl@s Bulletin*, vol. 6, no. 3, Nov. 2017, <http://docs.lib.purdue.edu/artlas/vol6/iss3/2>.

Lincoln, Matthew D. "Holding Out for a CV Hero: The Frick Computer Vision Symposium." *Matthew Lincoln, PhD*, 16 Apr. 2018, <https://matthewlincoln.net/2018/04/16/holding-out-for-a-cv-hero-the-frick-computer-vision-symposium.html>.

Linked Data - Design Issues. <https://www.w3.org/DesignIssues/LinkedData.html>. Accessed 17 Aug. 2017.

Linked Data API for JavaScript. Contribute to Linkeddata/RdfLib.js Development by Creating an Account on GitHub. 2011. Read-Write Linked Data, 2019. *GitHub*, <https://github.com/linkeddata/rdfLib.js>.

Linked Data Platform 1.0. <https://www.w3.org/TR/ldp/>. Accessed 5 Feb. 2019.

Linked Open Data | Yale Center for British Art. <https://britishart.yale.edu/collections/using-collections/technology/linked-open-data>. Accessed 22 Feb. 2019.

Linked Open Vocabularies (LOV). <https://lov.linkeddata.es/dataset/lov/vocabs/sim>. Accessed 10 Mar. 2019.

Linked Research. <https://linkedresearch.org/>. Accessed 28 May 2019.

LLC, ImageMagick Studio. "ImageMagick." *ImageMagick*, <https://imagemagick.org/>. Accessed 30 May 2019.

Loos, Ted. "'Photo Archives Are Sleeping Beauties.' Pharos Is Their Prince." *The New York Times*, 22 Dec. 2017. *NYTimes.com*, <https://www.nytimes.com/2017/03/14/arts/design/art-history-digital-archive-museums-pharos.html>.

LOUD: Linked Open Usable Data. <https://linked.art/loud/index.html>. Accessed 23 Feb. 2019.

Manovich, Lev. *Data Science and Digital Art History*. p. 26.

Marketakis, Yannis. *Dockerized Version of 3M*. 2017. 2019. *GitHub*, <https://github.com/ymark/3M-docker>.

Marmor, Max. "Art History and the Digital Humanities." *Zeitschrift Für Kunstgeschichte*, p. 5.

Medici Archive Project Mission | *The Medici Archive Project*. <http://www.medic.org/mission/>. Accessed 5 Feb. 2019.

Metaphactory. <https://metaphacts.com/product>. Accessed 23 Feb. 2019.

Mittal, Sudip, et al. "Thinking, Fast and Slow: Combining Vector Spaces and Knowledge Graphs." *ArXiv:1708.03310 [Cs]*, Aug. 2017. *arXiv.org*, <http://arxiv.org/abs/1708.03310>.

Mix'n'match. https://tools.wmflabs.org/mix-n-match/#/group/ig_art. Accessed 3 Mar. 2019.

Moretti, Franco. *Distant Reading*. Verso, 2013.

New York Art Resources Consortium |. <http://nyarc.org/>. Accessed 23 Feb. 2019.

Oldman, Dominic, et al. "Zen and the Art of Linked Data: New Strategies for a Semantic Web of Humanist Knowledge." *A New Companion to Digital Humanities*, edited by Susan Schreibman et al., John Wiley & Sons, Ltd, 2015, pp. 251–73. *Crossref*, doi:[10.1002/9781118680605.ch18](https://doi.org/10.1002/9781118680605.ch18).

Omeka. <https://omeka.org/>. Accessed 23 Feb. 2019.

"Ontotext GraphDBTM - a Semantic Graph Database Free Download." *Ontotext*, <https://www.ontotext.com/products/graphdb/>. Accessed 4 Mar. 2019.

Orphanides, K. G. "How Tim Berners-Lee's Inrupt Project Plans to Fix the Web." *Wired UK*, Feb. 2019. *www.wired.co.uk*, <https://www.wired.co.uk/article/inrupt-tim-berners-lee>.

Ota, Allyson. *Reconciling Smithsonian Library Data with VIAF*. Smithsonian Libraries, 8 Sept. 2016.

OWL - *Semantic Web Standards*. <https://www.w3.org/OWL/>. Accessed 3 Mar. 2019.

"Part III: Review of Funders of Digital Cultural Heritage Initiatives • CLIR." *CLIR*, <https://www.clir.org/pubs/reports/pub118/part3/>. Accessed 6 Feb. 2019.

Pastec, the Open Source Image Recognition Technology for Your Mobile Apps. <http://pastec.io/>. Accessed 10 Mar. 2019.

PHAROS: The International Consortium of Photo Archives. <http://pharosartresearch.org/>. Accessed 17 Mar. 2019.

Photo Archive (Getty Research Institute). <http://www.getty.edu/research/tools/photo/>. Accessed 23 Feb. 2019.

Photograph Archives | I Tatti | The Harvard University Center for Italian Renaissance Studies. <http://itatti.harvard.edu/berenson-library/collections/photograph-archives>. Accessed 24 Feb. 2019.

“Primo Library Resource Discovery Solution.” *Ex Libris*, <https://www.exlibrisgroup.com/products/primo-library-discovery/>. Accessed 23 Feb. 2019.

Project Blacklight. <http://projectblacklight.org/>. Accessed 8 Mar. 2019.

R: The R Project for Statistical Computing. <https://www.r-project.org/>. Accessed 23 Feb. 2019.

Rahman, Abrar. *Interactive Foreground Extraction with Superpixels.*

Rawson, Katie, and Trevor Muñoz. *Against Cleaning.* July 2016. [curatingmenus.org, http://www.curatingmenus.org/articles/against-cleaning/](http://www.curatingmenus.org/articles/against-cleaning/).

RDF - Semantic Web Standards. <https://www.w3.org/RDF/>. Accessed 23 Feb. 2019.

RDF Graph Literals and Named Graphs. <https://www.w3.org/2009/07/NamedGraph.html>. Accessed 9 Mar. 2019.

“Reasons to Share Your Data on Europeana Collections.” *Europeana Pro*, <https://pro.europeana.eu/page/reasons-to-share-your-data-on-europeana-collections>. Accessed 22 Feb. 2019.

ResearchSpace - a Digital Wunderkammer for the Cultural Heritage Knowledge Graph. <https://www.researchspace.org/>. Accessed 10 Sept. 2017.

Robert Sanderson. *EuropeanaTech Keynote: Shout It out LOUD*. <https://www.slideshare.net/azaro42/europeanatech-keynote-shout-it-out-loud>.

Rochkind, J. “Is the Semantic Web Still a Thing?” *Bibliographic Wilderness*, 28 Oct. 2014, <https://bibwild.wordpress.com/2014/10/28/is-the-semantic-web-still-a-thing/>.

Schermerhorn, L. W. *DIGITAL RESOURCES FOR THE HISTORY OF ART GRANT PROGRAM*. Kress Foundation.

ScienceMatters | *ScienceMatters*. <https://www.sciencematters.io/>. Accessed 17 Mar. 2019.

Searching Through Seeing: Optimizing Computer Vision Technology for the Arts | *The Frick Collection*. https://www.frick.org/interact/video/searching_seeing. Accessed 10 Mar. 2019.

Seguin, Benoit, et al. “Visual Link Retrieval in a Database of Paintings.” *Computer Vision – ECCV 2016 Workshops*, edited by Gang Hua and Hervé Jégou, Springer International Publishing, 2016, pp. 753–67.

Semantic Digital Publishing - Consortium for Open Research Data in the Humanities. <https://wiki.cordh.net/display/SDP/Semantic+Digital+Publishing>. Accessed 17 Mar. 2019.

Sheth, Amit, editor. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*. IGI Global, 2011. *Crossref*, doi:[10.4018/978-1-60960-593-3](https://doi.org/10.4018/978-1-60960-593-3).

SKOS Simple Knowledge Organization System Namespace Document - HTML Variant, 18 August 2009 Recommendation Edition. <https://www.w3.org/2009/08/skos-reference/skos.html>. Accessed 3 Mar. 2019.

Snapshot. <http://curatingmenus.org/articles/against-cleaning/>. Accessed 3 Oct. 2018.

Solid | Inrupt. <https://inrupt.com/solid>. Accessed 10 Mar. 2019.

Solr 5.5 Config Files for Florentine Drawings Project: Villaitatti/Florentine-Drawings-Solr-Config. 2016. Villa I Tatti | The Harvard University Center for Italian Renaissance Studies, 2018. *GitHub*, <https://github.com/villaitatti/florentine-drawings-solr-config>.

Super COOLSCAN 5000 ED from Nikon. <https://www.nikonusa.com/en/nikon-products/product-archive/film-scanners/super-coolscan-5000-ed.html>. Accessed 24 Feb. 2019.

Szegedy, Christian, et al. "Going Deeper with Convolutions." *ArXiv:1409.4842 [Cs]*, Sept. 2014. *arXiv.org*, <http://arxiv.org/abs/1409.4842>.

Szekely, Pedro, et al. "Connecting the Smithsonian American Art Museum to the Linked Data Cloud." *The Semantic Web: Semantics and Big Data*, edited by Philipp Cimiano et al., Springer Berlin Heidelberg, 2013, pp. 593–607.

Talbur, John R. "1 - Principles of Entity Resolution." *Entity Resolution and Information Quality*, edited by John R. Talbur, Morgan Kaufmann, 2011, pp. 1–37. *ScienceDirect*, doi:[10.1016/B978-0-12-381972-7.00001-4](https://doi.org/10.1016/B978-0-12-381972-7.00001-4).

Tan, W. R., et al. "Ceci n'est Pas Une Pipe: A Deep Convolutional Network for Fine-Art Paintings Classification." *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 3703–07. *IEEE Xplore*, doi:[10.1109/ICIP.2016.7533051](https://doi.org/10.1109/ICIP.2016.7533051).

Technical Standards for Digital Conversion of Text and Graphic Materials. Library of Congress, p. 28.

The CIDOC Conceptual Reference Model (CIDOC-CRM): PRIMER | CIDOC CRM. <http://www.cidoc-crm.org/lrmoo/Resources/the-cidoc-conceptual-reference-model-cidoc-crm-primer>. Accessed 3 Oct. 2018.

The Correlation between Semantic Visual Similarity and Ontology-Based Concept Similarity in Effective Web Image Search. 1st ed, Springer, 2012.

The Drawings of the Florentine Painters. <http://florentinedrawings.itatti.harvard.edu/>. Accessed 23 Feb. 2019.

The Linked Open Data Cloud. <https://lod-cloud.net/>. Accessed 5 Feb. 2019.

The Similarity Ontology. 16 Jan. 2013, <https://web.archive.org/web/20130116095414/http://kakapo.dcs.qmul.ac.uk/ontology/musim/0.2/musim.html>.

TinEye Reverse Image Search. <https://www.tineye.com/>. Accessed 10 Mar. 2019.

TMS Collections Collection Management Software, Museum Collections. <https://www.gallerysystems.com/products-and-services/tms-suite/tms/>. Accessed 8 Mar. 2019.

Toward Spatial Humanities : Historical GIS and Spatial History. Indiana University Press, 2014.

Travis, Charles. *Abstract Machine : Humanities GIS*. First edition., Esri Pres, 2015.

Tzompanaki, Katerina, and Martin Doerr. *Fundamental Categories and Relationships for Intuitive Querying CIDOC-CRM Based Repositories*. 2012, p. 153.

Usability - Develop the Usability Test. <http://www.utexas.edu/learn/usability/test.html>. Accessed 11 Apr. 2010.

Usability Toolkit. <http://www.stcsig.org/usability/resources/toolkit/toolkit.html>. Accessed 11 Apr. 2010.

User Interface Usability Evaluation with Web-Based Questionnaires. <http://oldwww.acm.org/perlman/question.html>. Accessed 11 Apr. 2010.

Verborgh, Ruben, and Miel Vander Sande. "The Semantic Web Identity Crisis: In Search of the Trivialities That Never Were." *Semantic Web*, no. pre-press, <http://www.semantic-web-journal.net/content/semantic-web-identity-crisis-search-trivialities-never-were>. Accessed 28 May 2019.

VIAF. <https://viaf.org/>. Accessed 3 Mar. 2019.

VISART IV. <https://visarts.eu/>. Accessed 10 Mar. 2019.

Web Technologies and Application: APWeb 2012 International Workshops: SENDE, IDP, MBC, Kunming, China, April 11, 2012. Proceedings. 1st ed, Springer, 2012.

Yang, Jun, et al. "Evaluating Bag-of-Visual-Words Representations in Scene Classification." *Proceedings of the International Workshop on Workshop on Multimedia*

Information Retrieval - MIR '07, ACM Press, 2007, p. 197. Crossref, doi: [10.1145/1290082.1290111](https://doi.org/10.1145/1290082.1290111).

Zorach, Diane. *Kress Foundation | Transitioning to a Digital World: Art History, Its Research Centers, and Digital Scholarship*. http://www.kressfoundation.org/research/transitioning_to_a_digital_world/. Accessed 18 Aug. 2017.

Zorich, Diane. *A Survey of Digital Cultural Heritage Initiatives and Their Sustainability Concerns*. Council on Library and Information Resources, 2003.

Zorich, Diane M. "Digital Art History: A Community Assessment." *Visual Resources*, vol. 29, no. 1–2, June 2013, pp. 14–21. *Taylor and Francis+NEJM*, doi: [10.1080/01973762.2013.761108](https://doi.org/10.1080/01973762.2013.761108).

Appendix A

```

import argparse
import cv2
import logging
import numpy as np
import os.path
import sys
import time

DEBUG = 0

class GrabCutExtraction:
    BG = 0
    FG = 1
    PR_FG = 3
    PR_BG = 2

    MAX_WIDTH = 480.0
    INIT_BG_STRIPE = 0.1

    def __init__(self, img, resize = False):
        self.in_img = img
        self.out_img = img.copy()
        self.resize = resize
        # Enforce max size for optimal processing time
        if resize:
            _, cols = self.out_img.shape[:2]
            factor = float(self.MAX_WIDTH)/float(cols)
            self.out_img = cv2.resize(self.out_img, (0,0), fx=factor,
fy=factor)
            # Blur to reduce noise impact
            self.out_img = cv2.GaussianBlur(self.out_img, (15, 15), 0)
            #Initialize
            self.mask = np.zeros(self.out_img.shape[:2], dtype = np.uint8)
# mask initialized to BG
            self.bgdmodel = np.zeros((1,65), np.float64)
            self.fgdmodel = np.zeros((1,65), np.float64)
            rows, cols = self.out_img.shape[:2]
            self.rect_init = (int(self.INIT_BG_STRIPE*cols),
int(self.INIT_BG_STRIPE*rows),

```

```

        int((1 - 2*self.INIT_BG_STRIPE)*cols), int((1 -
2*self.INIT_BG_STRIPE)*rows))

    def process(self):
        cv2.grabCut(self.out_img, self.mask, self.rect_init,
self.bgdmodel, self.fgdmodel, 1, cv2.GC_INIT_WITH_RECT)
        if DEBUG == 1:
            cv2.imshow('Initial mask', self.mask)
            cv2.waitKey()

    def morph_open(self, mask, kernel_size, erode_iter, dilate_iter):
        kernel = np.ones((kernel_size, kernel_size), np.uint8)
        eroded = cv2.erode(mask, kernel, iterations = erode_iter)
        opened = cv2.dilate(eroded, kernel, iterations = dilate_iter)
        return eroded, opened

    def refine(self):
        fg_mask = np.where((self.mask == self.FG) + (self.mask ==
self.PR_FG), 255, 0).astype('uint8')
        # Get rid of noise using morphological opening
        _, fg_mask = self.morph_open(fg_mask, 3, 3, 3)
        # Refine foreground/background mask
        eroded, opened = self.morph_open(fg_mask, 3, 8, 15)
        if DEBUG == 1:
            cv2.imshow('Eroded', eroded)
            cv2.imshow('Opened', opened)
            cv2.waitKey()
        has_fg = False
        for i in xrange(self.out_img.shape[0]):
            for j in xrange(self.out_img.shape[1]):
                if eroded[i, j] == 255:
                    self.mask[i, j] = self.FG
                    has_fg = True
                elif opened[i, j] == 255:
                    self.mask[i, j] = self.PR_BG
                else:
                    self.mask[i, j] = self.BG
        # Terminate if we haven't detected any big enough object.
        if has_fg:
            cv2.grabCut(self.out_img, self.mask, self.rect_init,
self.bgdmodel, self.fgdmodel, 1, cv2.GC_INIT_WITH_MASK)
        else:
            logging.error("No foreground object.")
        if self.resize:

```

```

        factor = float(self.in_img.shape[1])/
float(self.mask.shape[1])
        self.mask = cv2.resize(self.mask, (self.in_img.shape[1],
self.in_img.shape[0]), interpolation = cv2.INTER_NEAREST)

    def fill_holes(self, bin_img):
        im_floodfill = bin_img.copy()
        # Mask used to flood filling.
        # Notice the size needs to be 2 pixels than the image.
        h, w = bin_img.shape[:2]
        fill_mask = np.zeros((h+2, w+2), np.uint8)

        # Floodfill from point (0, 0)
        cv2.floodFill(im_floodfill, fill_mask, (0,0), 255);

        # Invert floodfilled image
        im_floodfill_inv = cv2.bitwise_not(im_floodfill)

        # Combine the two images to get the foreground.
        out = bin_img | im_floodfill_inv
        if DEBUG == 1:
            cv2.imshow('Mask', out)
            cv2.waitKey()
        return out

    def write_output(self, directory, file_name):
        show_mask = np.where((self.mask == self.FG) + (self.mask ==
self.PR_FG), 255, 0).astype('uint8')
        # Smooth shape using morphological opening
        _, show_mask = self.morph_open(show_mask, 5, 20, 20)
        show_mask = self.fill_holes(show_mask)
        output = np.zeros(self.in_img.shape, np.uint8)
        output = cv2.bitwise_and(self.in_img, self.in_img,
mask=show_mask)
        cv2.imwrite(os.path.join(directory, file_name), output)
        if DEBUG == 1:
            cv2.imshow('Output', output)
            cv2.waitKey()

if __name__ == '__main__':
    # Configure logging
    logging.basicConfig(level = logging.INFO)
    # Construct the argument parser and parse the arguments
    ap = argparse.ArgumentParser(description="Extract object from
image.")

```

```

    ap.add_argument("-i", "--input", required = True,
        help = "Path to the directory with images")
    ap.add_argument("-o", "--output", default = "results", required =
False,
        help = "Path to the directory where output images should be
placed")
    args = vars(ap.parse_args())
    in_dir = os.path.abspath(args["input"])
    out_dir = os.path.abspath(args["output"])

    # Output directory should be created by the program
    if os.path.isdir(out_dir):
        logging.critical("Output directory %s already exists. Please
(re)move it before proceeding.", out_dir)
        sys.exit()
    else:
        os.makedirs(out_dir)

    # Process image files in input directory
    in_files = [os.path.join(in_dir, f) for f in os.listdir(in_dir)
if os.path.isfile(os.path.join(in_dir, f))]
    for fn in in_files:
        logging.info("Processing %s...", os.path.basename(fn))
        img = cv2.imread(fn)
        if img is None:
            logging.error("Image %s could not be read.", fn)
            continue
        start = time.time()
        extraction = GrabCutExtraction(img, resize = True)
        extraction.process()
        extraction.refine()
        extraction.write_output(out_dir, os.path.basename(fn))
        end = time.time()
        logging.debug("Execution time: %d", end - start)

```

Appendix B

```

import cv2
import numpy as np
from matplotlib import pyplot as plt

import os
import sys
import math
import csv

class physicDim:

    def __init__(self, calc_width=1000, bShow=True):
        self.calc_width = calc_width
        self.bShow = bShow

    def proc(self, img_path):

        base_img = cv2.imread(img_path)

        h, w = base_img.shape[:2]
        new_h, new_w = int(self.calc_width / w * h),
self.calc_width

        # gray and resizing
        resize = cv2.resize(base_img, (new_w, new_h))
        gray = cv2.cvtColor(resize, cv2.COLOR_BGR2GRAY)
        # cv2.imshow("gray", gray)

        # crop the image to isolate the black hole(0)
        if new_h > new_w:
            trans = np.transpose(gray)
            trans = cv2.GaussianBlur(trans, (11, 11), 0)
            for right in range(int(new_w*3/4), new_w):
                if np.amin(trans[right][int(new_h/4):int(3*new_h/
4)]) == 0:
                    break
            for left in range(int(new_w/4), 0, -1):
                if np.amin(trans[left][int(new_h/4):int(3*new_h/
4)]) == 0:

```

```

        break
    if right == left:
        left, right = 0, new_w
        top, bottom = 0, new_h
    else:
        trans = gray
        trans = cv2.GaussianBlur(trans, (11, 11), 0)
        for top in range(int(new_h/4), 0, -1):
            if np.amin(trans[top][int(new_w/4):int(3*new_w/4)])
== 0:
                break
            for bottom in range(int(new_h*3/4), new_h):
                if np.amin(trans[bottom][int(new_w/4):int(3*new_w/
4)]) == 0:
                    break
                if top == bottom:
                    top, bottom = 0, new_h
                    left, right = 0, new_w

        top, bottom = 0, new_h
        left, right = 0, new_w

        crop = gray[top:bottom, left:right]

        # remove noising
        noise = cv2.fastNlMeansDenoising(crop, None, 20, 7, 21)

        # CLAH algorithm(Contrast Limited Adaptive Histogram
Equalization)
        clahe = cv2.createCLAHE(clipLimit=2.0, tileGridSize=(8, 8))
        cla = clahe.apply(noise)

        # otsu's thresholding
        gaus_blur = cv2.GaussianBlur(cla, (5, 5), 0)
        otsu_thresh = cv2.threshold(gaus_blur, 0, 255,
cv2.THRESH_BINARY + cv2.THRESH_OTSU)[1]

        kernel_1 = np.ones((5, 5), np.uint8)
        kernel_2 = np.ones((7, 7), np.uint8)
        otsu_thresh = cv2.dilate(otsu_thresh, kernel_1,
iterations=1)
        otsu_thresh = cv2.erode(otsu_thresh, kernel_2,
iterations=1)

        # horizontal line removing

```

```

kernel_h = np.ones((1, 15), dtype=int)
erode_h = cv2.erode(otsu_thresh, kernel_h)
dilate_h = cv2.dilate(erode_h, kernel_h)

# vertical line removing
kernel_v = np.ones((15, 1), dtype=int)
erode_v = cv2.erode(otsu_thresh, kernel_v)
dilate_v = cv2.dilate(erode_v, kernel_v)

# bitwise vertical line and horizontal line
remove = cv2.bitwise_and(dilate_h, dilate_v)

# largest contour extraction
max_area = 0.0
max_idx = None
_, contours, hierarchy = cv2.findContours(remove,
cv2.RETR_TREE, cv2.CHAIN_APPROX_SIMPLE)
for idx in range(len(contours)):
    area = cv2.contourArea(contours[idx])
    if max_area < area:
        max_area = area
        max_idx = idx

c_h, c_w = crop.shape[:2]
UpLeft = (0, 0)
DownLeft = (0, c_h)
UpRight = (c_w, 0)
DownRight = (c_w, c_h)

min_upLeft, min_upRight, min_downLeft, min_downRight = c_h
+ c_w, c_h + c_w, c_h + c_w, c_h + c_w
corner_upLeft, corner_upRighth, corner_downLeft,
corner_downRight = None, None, None, None
# extract the four points of corner

show = resize[top:bottom, left:right]
for pt in contours[max_idx]:

    dis_upLeft = self.distance(pt[0], UpLeft)
    dis_upRight = self.distance(pt[0], UpRight)
    dis_downLeft = self.distance(pt[0], DownLeft)
    dis_downRight = self.distance(pt[0], DownRight)

# find the upleft corner
if min_upLeft > dis_upLeft:

```



```

        min_upLeft = dis_upLeft
        corner_upLeft = pt[0]

    if min_upRight > dis_upRight:
        min_upRight = dis_upRight
        corner_upRigth = pt[0]

    if min_downLeft > dis_downLeft:
        min_downLeft = dis_downLeft
        corner_downLeft = pt[0]

    if min_downRight > dis_downRight:
        min_downRight = dis_downRight
        corner_downRight = pt[0]

    if self.bShow:
        # show = resize[top:bottom, left:right]

        cv2.drawContours(show, contours, -1, (0, 255, 0), 1)
        cv2.drawContours(show, contours, max_idx, (0, 0, 255),
2)

        cv2.circle(show, (corner_upLeft[0], corner_upLeft[1]),
3, (0, 255, 255), 2)
        cv2.circle(show, (corner_upRigth[0],
corner_upRigth[1]), 3, (0, 255, 255), 2)
        cv2.circle(show, (corner_downLeft[0],
corner_downLeft[1]), 3, (0, 255, 255), 2)
        cv2.circle(show, (corner_downRight[0],
corner_downRight[1]), 3, (0, 255, 255), 2)

        result = cv2.resize(resize, (int(resize.shape[1]/2),
int(resize.shape[0]/2)))
        cv2.imshow("result", result)

        key = cv2.waitKey(500)
        if key == ord('i'): # ignore
            return "ignore"
        elif key == ord('o'): # okay
            pass
        elif key == ord('n'): # next
            return "pass"

# calculate the real corner from resized to baseImage
ul = (corner_upLeft[0] + left, corner_upLeft[1] + top)

```

```

ur = (corner_upRigth[0] + left, corner_upRigth[1] + top)
dl = (corner_downLeft[0] + left, corner_downLeft[1] + top)
dr = (corner_downRight[0] + left, corner_downRight[1] +
top)

re_ul = (ul[0] * w / new_w, ul[1] * w / new_w)
re_ur = (ur[0] * w / new_w, ur[1] * w / new_w)
re_dl = (dl[0] * w / new_w, dl[1] * w / new_w)
re_dr = (dr[0] * w / new_w, dr[1] * w / new_w)

x_, y_, w_, h_ = cv2.boundingRect(contours[max_idx])
c_left, c_top, c_width, c_height = x_ + left, y_ + top, w_,
h_

percent_left = c_left * 100 / new_w
percent_top = c_top * 100 / new_h
percent_width = c_width * 100 / new_w
percent_height = c_height * 100 / new_h

# calculate the angle
angle = 0.0
angle += math.atan(float(ul[1] - ur[1]) / float(ul[0] -
ur[0]))
angle -= math.atan(float(dl[0] - ul[0]) / float(dl[1] -
ul[1]))
angle += math.atan(float(dr[1] - dl[1]) / float(dr[0] -
dl[0]))
angle -= math.atan(float(ur[0] - dr[0]) / float(ur[1] -
dr[1]))

angle = (angle / 4.0) * (180 / math.pi)

fpath, fname = os.path.split(img_path)
fname, ext = os.path.splitext(fname)

dict = {}
dict["fname"] = fname
dict["fpath"] = fpath
dict["upperLeft"] = re_ul
dict["upperRight"] = re_ur
dict["downLeft"] = re_dl
dict["downRight"] = re_dr
dict["angle"] = angle

dict["percentLeft"] = percent_left
dict["percentTop"] = percent_top

```

```

    dict["percentWidth"] = percent_width
    dict["percentHeight"] = percent_height
    return dict

    @staticmethod
    def distance(point1, point2):
        return math.sqrt((point1[0] - point2[0]) ** 2 + (point1[1]
- point2[1]) ** 2)

def scan(folder):

    list = []
    for f in os.listdir(folder):
        path = os.path.join(folder, f)

        # scan only .jpg image files
        if os.path.isfile(path) and os.path.splitext(path)[1] ==
'.jpg':
            list.append(path)
            if os.path.isdir(path):
                list.extend(scan(path + "/"))

    return list

"""
    csv file format

+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+
| file name| UpperLeft|UpperRight| DownLeft| DownRight|      Angle|
%fromLeft| % fromTop|  % Widht | % Height |
+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+
|          ...|      ...|      ...|      ...|      ...|      ...|
...|          ...|      ...|      ...|      ...|      ...|
|          ...|      ...|      ...|      ...|      ...|      ...|
...|          ...|      ...|      ...|
"""

if __name__ == '__main__':

    phy = physicDim(calc_width=1000, bShow=False)

```

```

work_dir = './Measure Samples-20170731T092636Z-001/'
if os.path.isdir(work_dir):
    sys.stdout.write("working directory: {}
\n".format(work_dir))
    file_list = scan(work_dir)
else:
    sys.stdout.write("No such directory, {}
\n".format(work_dir))

# front : "recto", back : "verso"
csv_path = os.path.join(work_dir, 'result.csv')
sys.stdout.write("result csv file: {}\n".format(csv_path))

# write the result as a csv file
with open(csv_path, 'w') as csvfile:

    # write the head of file
    head_str = "FileName, FilePath, UpperLeft, UpperRight,
DownLeft, DownRight, Angle, % fromLeft, % fromTop, % Widht, %
Height\n"
    csvfile.write(head_str)

    for f in file_list[:]:

        line = ""
        if f.find("recto") != -1:
            recto_path = f
            verso_path = f.replace("recto", "verso")
            sys.stdout.write("{}\n".format(verso_path))

            if os.path.isfile(recto_path) and
os.path.isfile(verso_path):

                # main proc
                res = phy.proc(verso_path)
                if isinstance(res, dict):
                    line = line + "{},".format(res["fname"])
                    line = line + "{},".format(res["fpath"])
                    line = line + "({:7.3f} {:
7.3f}),".format(res["upperLeft"][0], res["upperLeft"][1])
                    line = line + "({:7.3f} {:
7.3f}),".format(res["upperRight"][0], res["upperRight"][1])
                    line = line + "({:7.3f} {:
7.3f}),".format(res["downLeft"][0], res["upperLeft"][1])

```

```

        line = line + "({:7.3f} {:
7.3f}),".format(res["downRight"][0], res["downRight"][1])
        line = line + "{:
7.4f},".format(res["angle"])

        line = line + "{:7.1f}
%,".format(res["percentLeft"])
        line = line + "{:7.1f}
%,".format(res["percentTop"])
        line = line + "{:7.1f}
%,".format(res["percentWidth"])
        line = line + "{:7.1f}%
\n".format(res["percentHeight"])

        csvfile.write(line)

    elif isinstance(res, str):
        if res == "pass":
            sys.stdout.write("PASSEd {}
\n".format(os.path.basename(verso_path)))
            continue
        elif res == "ignore":
            sys.stdout.write("IGNORED {}
\n".format(os.path.basename(verso_path)))
            continue

```

Estratto per riassunto della tesi di dottorato

Studente: Lukas Klic

Matricola: 956256

Dottorato: Informatica

Ciclo: XXXI

Titolo della tesi: Digital Publishing and Research Infrastructure: an Institutional Roadmap

Abstract

The burgeoning field of Digital Humanities has seen a great deal of interest in methodologies that support the exploration, cross-pollination, and programmatic analysis of heritage collections across the web of data. Although the heritage community has generally agreed that these data should be semantically enriched using the CIDOC Conceptual Reference Model and published as Linked Open Data, a lack of agreement at both data and infrastructural levels has hindered advancements that would allow for greater data integration and computational exploration. This project provides an institutional roadmap for publishing such data in a Semantic Web research environment, proposing a set of best practices for the community. Using a collection of 230,000 images and index metadata, this project presents methodologies and tools for data cleaning, reconciliation, enrichment, and transformation for publishing in a native Resource Description Framework system. A semantic framework for integrating computer vision services enables subsequent enrichment and visual analysis, enabling the mass-digitization of heritage collections with minimal burden on institutions, all while ensuring the long-term preservation and interoperability of these data at a global scale.