SPRING 2020

Savannah Lake

Master of Library and Information Studies Candidate Portfolio

Advisor: Jean-François Blanchette
savannahlake.github.io/mlisportfolio

Table of Contents

About Me

Information is only usable if it is retrievable. I am interested in how to build user-centric collections, repositories, and systems that promote discoverability and information literacy. As such, my master's studies have focused on metadata, information architecture, and description.

In addition to my studies, I have also sought opportunities that would expose me to diverse use cases and information types. Institutions I was fortunate to work at include an academic library, museum, community archive, and research institute, where I developed a deeper understanding of information-seeking behavior, information management, and practical worflows.

This portfolio showcases my academic and professional achievements, offering a holistic sense of how I would contribute to the library and information science field.

Academics

My academic studies at UCLA focused on metadata and information architecture, with the ultimate goal of learning how to leverage these concepts in information and collections management. As such, I sought courses that would build my understanding of foundational and theoretical concepts as well as courses that would build my technical skills.

This section features a list of classes I have taken during my time at UCLA, as well as examples of my coursework, including my major paper, elective coursework, core coursework.

Courses Taken

Fall 2018
Artifacts and Cultures, Johanna Drucker (IS 211)
Description and Access, Gregory Leazer (IS 260)
Archives, Records, and Memory, Anne Gilliland (IS 431)

Winter 2019
Data Management and Practice, Jillian Wallis (IS 262A)
Systems and Infrastructures, Jean-François Blanchette (IS 270)
Human-Computer Interaction, Leah Lievrouw (IS 272)

Spring 2019
Data Curation and Policy, Jillian Wallis (IS 262B)
Metadata, Jonathan Furner (IS 464)
Values and Communities, Sarah Roberts (IS 212)
Content Management Systems, Gregg Rugolo (IS 289)

Fall 2019
Digital Humanities, Miriam Posner (DH 201)
Computer Systems and Programming, Joshua Gomez (IS 271)
Internship, Getty Research Institute, Jean-François Blanchette, Melissa Gill (IS 498)

Winter 2020
History of Books and Literacy Technologies, Johanna Drucker (IS 202)
Historical Research Methods, Shawn VanCour (IS 281)
Descriptive Cataloging, Luiz Mendes (IS 461)

Spring 2020
Digital Asset Management, Linda Tadic (IS 289)
User Experience Research, Mary Zide (IS 279)
Subject Cataloging, Luiz Mendes (IS 462)
Internship, Getty Research Institute, Jean-François Blanchette, Patrica Harpring (IS 498)

Major Paper

Metadata // IS 464
Jonathan Furner
Spring 2019

Abstract
The Fowler Museum is a global arts and cultures museum, specializing in art from Africa, Asia, Indigenous North and South Americas, and the Pacific. Their collections span over two millennia, and include over 120,000 art and ethnographic objects as well as 600,000 archaeological objects, 20,000 textiles, and 400 silver works. The diversity and extent of their holdings could attract a lot of users from across the globe, given the holdings' origins. Despite this, the Fowler's digital collections are quite minimal, comprising of just 2,007 objects and lacking encoded, consistent, and robust metadata. A more strategic approach to metadata could enable the Fowler to better serve its users and the community at large—particularly in the domains of research and repatriation. This comprehensive metadata strategy addresses both of these use cases, offering recommendations on metadata schemata, vocabularies, rights metadata, and logistics to best optimize the digital collections' metadata.

Inclusion in portfolio
I am deeply interested in metadata and its relation to access. This paper allowed me to think critically about these issues, grounding my analysis in an institution and two use cases. Because of this framework, I was able to offer concrete recommendations for a path forward. This is the sort of work that I hope to do with my MLIS degree, making it an fitting major paper.

The Fowler Museum's Digital Collections: A Metadata Strategy for Facilitating Research and Repatriation

Savannah Lake

Information Studies 464: Metadata

Professor Jonathan Furner

June 7, 2019

The Fowler Museum at the University of California, Los Angeles is a global arts and cultures museum, specializing in art from Africa, Asia, Indigenous North and South Americas, and the Pacific. Their collections span over two millennia, and include over 120,000 art and ethnographic objects as well as 600,000 archaeological objects, 20,000 textiles, and 400 silver works.[1] The diversity and extent of their holdings could attract a lot of users from across the globe, given the holdings' origins. Despite this, the Fowler's digital collections are quite minimal, comprising of just 2,007 objects and lacking encoded, consistent, and robust metadata.[2] A more strategic approach to metadata could enable the Fowler to better serve its users and the community at large—particularly in the domains of research and repatriation. The following comprehensive metadata strategy will address both of these use cases, offering recommendations on metadata schemata, vocabularies, rights metadata, and logistics to best optimize the digital collections' metadata.

The Fowler Museum's Current Metadata Practices

While the Fowler's website is attractive and robust, complete with resources like school curricula, audio guided tours, and collection-related videos, the digital collections themselves are somewhat lacking. Of their over 740,000 objects, only 2,007 are online. Additionally, approximately 650 objects are on view at the museum.[3] Given the large overlap between the digital collections and the objects on view at the museum, a substantial portion of the Fowler's holdings are not discoverable. In addition to the minimal representation of the collection online, the use of metadata is not built to promote discovery, context, aggregation, or reuse.

Metadata Issues

Both the quality of the Fowler's collections metadata and its technical implementation impact the usability and efficacy of their digital collections. With regard to metadata quality, the collection faces problems with consistency and detail. While most entries include the object name, place of origin, cultural group, materials used, dimensions, credit line, and accession

---

[1] "Collections Overview," Fowler Museum, https://www.fowler.ucla.edu/collections/home/. (Accessed May 31, 2019).
[2] "Products Archive," Fowler Museum, https://www.fowler.ucla.edu/collections/. (Accessed May 31, 2019).
[3] Suzanne Muchnic, 2013, "UCLA's Fowler Museum Turns 50 in Worldly Fashion," *Los Angeles Times*, https://www.latimes.com/entertainment/arts/la-xpm-2013-sep-28-la-et-cm-ucla-fowler-museum-50-20130929-story.html.

number,[4] other entries include date and artist.[5] Some omit cultural group,[6] while others share information as to whether the piece is currently on display in the museum.[7] This lack of consistency compromises the collection's searchability as well as the object descriptions, making it harder to fully access and understand the collection.

Further, some of the items are given further description and context through supplemental pages, including the Andean ceramics and the Lega figures,[8] for example. However, these pages are not linked back to individual object catalog records. Accordingly, if you were not browsing the site at large, you would not necessarily gain this contextual information. Valuable metadata, like dates, that are present in these supplemental pages are at times missing from individual object pages, as seen with the Lega figure overview[9] and a spoon from the collection.[10]

Further impeding access, discovery, and retrieval of these items is the fact that item metadata are not encoded in the back-end with tags corresponding to a certain schema. This suggests that a standardized schema is not being employed. Instead, object metadata are denoted in the HTML through simple paragraph breaks. While current search engines prioritize page text and linking patterns within their algorithms, the lack of metadata HTML tags does represent a loss in retrieval and access.[11] Further, the lack of a standardized and encoded schema hinders the collection's ability to be aggregated or integrated into other repositories or works. A researcher, for example, would not be able to run a script and scrape the metadata as easily, preventing reuse of the collection.

Finally, reuse of the collection would be contingent on understanding the permission rights surrounding these items. Especially as the Fowler's digital collections include skulls and Indigenous art, clear provisions are needing for marking what can be used in what capacity and

---

[4] "X2007.21.90 Lega Spoon," Fowler Museum, https://www.fowler.ucla.edu/product/x2007-21-90-lega-spoon/. (Accessed May 31, 2019).
[5] "X92.311 Lambayeque Vessel," Fowler Museum, https://www.fowler.ucla.edu/product/x92-311-lambayeque-vessel/. (Accessed May 31, 2019).
[6] "X95.38.207a,b Betel Mortar," Fowler Museum, https://www.fowler.ucla.edu/product/x95-38-207ab-betel-mortar/. (Accessed May 31, 2019).
[7] "X91.410 Drinking Horn," Fowler Museum, https://www.fowler.ucla.edu/product/x91-410-drinking-horn/. (Accessed May 31, 2019).
[8] "Lega Figures," Fowler Museum, https://www.fowler.ucla.edu/collections/lega-figures/. (Accessed May 31, 2019).
[9] "Lega Figures," Fowler Museum.
[10] "X2007.21.90 Lega Spoon," Fowler Museum.
[11] Jenn Riley, 2010, "Glossary of Metadata Standards," http://jennriley.com/metadatamap/seeingstandards_glossary_pamphlet.pdf, 4.

context. Administrative and rights metadata would facilitate user interaction with and understanding of the collection.

Open Graph

While the Fowler's digital collections do not utilize encoded metadata to promote discovery, access, aggregation, and reuse, they do encode Open Graph tags to promote sharing over social media. Open Graph is a protocol that "enables any web page to become a rich object in a social graph" so that web pages enjoy the same functionality as other objects on social media.[12] Essentially, the protocol offers a set of tags that allow website developers to control what is displayed on social media platforms when users link to these web pages. Originally created by Facebook and now maintained by the Open Web Foundation, the protocol works on major social networks, including Facebook, Twitter, and LinkedIn.

While Open Graph is not intended to optimize a site for search engines, search algorithms likely account for Open Graph data, given the prominence of social networks within the Internet ecosystem.[13] Nevertheless, Open Graph is not a sufficient substitute for a metadata standard. On the Fowler's website, for example, Open Graph tags are limited to the URL, site name, image dimensions, site locale, title, and description—the latter of which serves as an unstructured, catch-all category for the object's descriptive metadata. This set-up fails to provide sufficient context for the discovery and understanding of records, favoring social media presence and sharing over resource description and retrieval.

This oversight is perhaps because the Fowler's website was created by Citrus Studios, a branding and digital agency. When marketing its web development services, Citrus Studios calls attention to their abilities in responsive web design, user experience, and branding—as opposed to metadata or searching functionalities.[14] Accordingly, it is possible they are not information specialists with a strong understanding of information management and stewardship.

Potential Users of the Fowler's Digital Collections Metadata

[12] "Open Graph Protocol," http://ogp.me/. (Accessed May 31, 2019).
[13] "Open Graph and Its Impact on SEO," Yakaferci, http://www.yakaferci.io/open-graph/. (Accessed May 31, 2019).
[14] "Responsive Web Design and Development Services," Citrus Studios, https://www.citrusstudios.com/online-marketing-services/responsive-web-design-development/. (Accessed June 1, 2019).

The Fowler's collections are particularly apt for two use cases that this proposed metadata strategy will cover: research and repatriation. With regard to research, the Fowler was founded in 1963 to complement the archaeology, anthropology, and ethnography programs on campus. The collection reflects its interdisciplinary beginnings, encompassing fields such as art history, architecture, anthropology, and archaeology in a way that many other museums do not. Currently, the museum is affiliated with the UCLA School of Arts and Architecture, with departmental classes like "World Arts and Cultures 24: World Arts, Local Lives" focused entirely on researching and understanding the Fowler's collections.[15] Given this, a strong user base of the collection materials are researchers from various fields.

The second user base that would benefit from the proposed metadata strategy would be those interested and involved in repatriation. The Fowler contains many Indigenous artworks, religious pieces, and even human remains as part of its archaeology efforts—all of which are contenders for repatriation. Currently, the Fowler has made several efforts to repatriate funerary objects and remains, complying with the Native American Graves Protection and Repatriation Act (NAGPRA) of 1990. These regulations require federally funded institutions to return cultural items such as human remains, funerary objects, sacred objects, and objects of cultural patrimony to Native American and Native Hawaiian organizations.[16] Per the *Federal Register*, the Fowler has submitted 13 notices of repatriation in accordance with NAGPRA, primarily for funerary objects and remains.[17] There remain Indigenous sacred objects and objects of cultural patrimony in their collection, where the Fowler's repatriation efforts might extend.

Metadata could support such repatriation efforts, particularly with regard to digital repatriation. Currently the Fowler is physically repatriating items. A more robust metadata and digital strategy could allow for digital repatriation or a post-custodial model for some items. Digital repatriation has certain attributes that require careful consideration before implementing. While it does allow for low-cost surrogates of materials to be returned to communities—or in a post-custodial model, to be retained by a memory institution—the "ease with which [digital resources] can be copied, distributed, and revised; their ability to exist in multiple locations at

---

[15] "World Arts and Cultures Courses," UCLA General Catalog 2018-19, https://catalog.registrar.ucla.edu/ucla-catalog18-19-1398.html. (Accessed June 3, 2019).
[16] "Archaeology," Fowler Museum, https://www.fowler.ucla.edu/archaeology/. (Accessed June 1, 2019).
[17] "Document Search Results for '"Notice of Intent To Repatriate Cultural Items" Fowler,'" Federal Register, https://www.federalregister.gov/documents/search?conditions%5Bterm%5D=%22Notice+of+Intent+To+Repatriate+Cultural+Items%22+fowler. (Accessed June 2, 2019).

once; and their ephemeral nature" requires robust metadata in order to ensure the long-term care and preservation of materials.[18]

Proposed Metadata Schema

While the library and archival worlds have developed and used metadata schemata like MARC, Encoded Archival Description (EAD), Describing Archives: A Content Standard (DACS), and Metadata Encoding and Transmission Standard (METS) since the 1970s, the museum world was slower to catch on to standardized metadata schemata. Fundamentally, the orientation of museums is to attract visitors through unique holdings, which means they are less agreeable to consensus and collaboration with "competitor" institutions. This has changed, however, since the late 1990s. Given the benefits of shared cataloging, particularly in the digital space, museums have begun to implement standardized metadata schemata, including Categories for the Description of Works of Art (CDWA), Visual Resources Association Core (VRA Core), Cataloging Cultural Objects (CCO), and Dublin Core.[19]

A Survey of Metadata Schema Options

While all of these schemata have strengths, some are less suited to the Fowler's collection and institutional practices. Currently, given the minimal number of objects posted online and the lack of robust metadata practices, there seem to be financial constraints preventing the Fowler from creating a more accessible, usable digital collection. Adopting CDWA as a metadata schema, then, would not be an apt fit: while thorough, CDWA's 540 categories and subcategories for description as well as their correlating authority files would require a substantial investment in time, labor, and maintenance that does not align with the Fowler's limited capacity for information management.[20]

VRA Core represents an interesting opportunity for the Fowler, as its strength come from its ability to describe both an object and its digital surrogate. Creating and clearly documenting

---

[18] Kimberly Christen, 2011, "Opening Archives: Respectful Repatriation," *The American Archivist* 74 (Spring/Summer 2011), 187.
[19] Anne Gilliland, 2016. "Setting the Stage" In *Introduction to Metadata*, by Murtha Baca. http://www.getty.edu/publications/intrometadata.
[20] "Categories for the Description of Works of Art (CDWA)," Getty Research Institute, http://www.getty.edu/research/publications/electronic_publications/cdwa/introduction.html#general. (Accessed June 2, 2019).

this boundary could prove helpful, especially in thinking about digital repatriation. However, VRA Core is chiefly intended to capture work records that can be affiliated with multiple image records.[21] That is not necessarily the case with the Fowler's collections, which primarily have one image for each work. Further, similar to CDWA, there is a level of complexity intrinsic to VRA Core that would require the Fowler to make a strong investment in information management. Not only does VRA Core require description for works and their digital surrogates, it also requires descriptions of collections—and creating relationships between all three of these elements. While these features allow for a more robust information experience, it may be outside the financial means and scope of the Fowler.

CCO is similarly robust. In addition to a metadata element set that can map onto VRA Core, CDWA, Dublin Core, and MARC, CCO provides extensive guidelines on formatting data, authorities, and controlled vocabularies.[22] While several of their required categories are a great fit for the Fowler's collections—including "current location," within the context of repatriation—some make less sense. Requiring an authority for the "controlled creator" field, for example, does not necessarily make sense for the Fowler, given the unknown or ambiguous creator for many of their objects, as well as the inadequacy of many vocabularies in representing non-Western artists. While use of controlled vocabularies is encouraged and will be detailed later on in this report, metadata schemata should have some flexibility, and understand that not all objects will be represented within a controlled vocabulary. Further, CCO requires a controlled vocabulary for the "controlled subjects" field. This would involve a substantial financial investment from the Fowler, as topical, subject metadata are wholly missing from the current online collection.[23]

Ultimately, the key is making the Fowler's collections searchable, findable, contextualized, and shareable—all while respecting the reality of the Fowler's budgetary and staffing limitations. Dublin Core, then, is the best fit, as it is a simple, low-cost metadata standard for digital objects. The schema has fifteen core metadata fields, all of which are both optional and repeatable. Accordingly, Dublin Core is meant to be "extremely simple, flexible,

[21] "VRA Core 4.0 Introduction." 2014. http://www.loc.gov/standards/vracore/VRA_Core4_Intro.pdf.
[22] Baca et. al., 2006, "Cataloging Cultural Objects: A Guide to Describing Cultural Works and Their Images," American Library Association, http://cco.vrafoundation.org/index.php/toolkit/cco_pdf_version/, 1.
[23] Baca et. al., "Cataloging Cultural Objects," 44-45.

and extensible" in order to encourage wide adoption online.[24] Despite this simplicity, Dublin Core provides enough fields to support description, retrieval, and preservation of digital collections.

Dublin Core and the Research User Base: Improving Access, Search, and Reuse

Researchers using the digital collections are served by the choice of Dublin Core as a metadata schema for the collection in terms of access. From a logistical standpoint, of all the schemata Dublin Core encourages the fastest upload of records as possible given the simplicity and straightforwardness of the schema. This is especially important as the vast majority of the Fowler's holdings are not available online or physically on view at the museum. Adding these records would open up research opportunities for scholars from a number of fields.

Utilizing Dublin Core as a metadata schema, as opposed to the collection's current homegrown schema, also allows for easier integration with image repositories like Artstor and the Online Archive of California.[25, 26] Dublin Core is the baseline format required for resources shared via the Open Archives Initiative Protocol for Metadata Harvesting, a technology that automates metadata sharing to enable cross-repository discovery.[27] Given the museum world's relatively late adoption of standardized metadata schemas, these cross-repository initiatives have had limited success, sometimes suffering from incomplete metadata records. [28] All the same, it is a resource that scholars can turn to—and a resource that will only improve with the continued commitment to responsible information management practices and contributions from museums.

The "rights" field is perhaps Dublin Core's strongest addition to the current Fowler record. For researchers, this field could include a rights statement (or a URL directing users to a rights statement) regarding the reuse of images. Rights metadata—for objects and their digital surrogates—ensure compliance with intellectual property laws, give researchers clarity and

---

[24] Stephen J. Miller, 2011, *Metadata for Digital Collections: A How-to-Do-It Manual*, London, United Kingdom: Facet Publishing, 51.
[25] "Technical Overview," Online Archive of California (OAC) / Calisphere Contributor Help Center, https://help.oac.cdlib.org/support/solutions/articles/9000081989-technical-overview. (Accessed June 4, 2019).
[26] "Metadata Policy," Artstor, https://www.artstor.org/contribute/metadata-policy/. (Accessed June 4, 2019).
[27] Riley, "Glossary of Metadata Standards," 11.
[28] Gilliland, "Setting the Stage" In *Introduction to Metadata*.

security about reuse, and set the foundation for the collection as a whole to be easily integrated into repositories like Artstor.[29]

Dublin Core and Repatriation: Enabling Collaboration and Metadata Justice

Dublin Core can serve repatriation efforts, as it facilitates digital technologies that "harness the collaborative potential between collecting institutions and indigenous communities."[30] One platform utilizing metadata to facilitate respectful digital repatriation is the Mukurtu content management system. Mukurtu is an open-source software that allows Indigenous communities to create and manage online collections of cultural heritage. Initially created through a collaboration with the Warumungu peoples of Australia and archivist Kimberly Christen of Washington State University (WSU), the platform introduces metadata that expresses Warumungu knowledge sets missing from Dublin Core through fields like "cultural narrative" and "traditional knowledge."[31] Accordingly, within Mukurtu each object can have multiple records, featuring the Mukurtu metadata that respects Indigenous knowledge while also maintaining institutional records and all of the history they bear. The Plateau Peoples' Web Portal, for example, uses Mukurtu. Many of their items are housed at WSU, which uses Dublin Core. Accordingly, for an object entry, Dublin Core records from WSU are shown as "institutional catalogue records" alongside records created by the community shown as "tribal catalogue records."[32] This juxtaposition gives both records equal footing, correcting the centuries' long bias and imposition of Western knowledge organization on Indigenous cultural materials. It also allows users to see the differences, additions, and corrections the tribal records have regarding the institutional record, revealing how "history is indeed made, unmade, and negotiated over time" and calling into question the primacy and orientation of institutional records.[33]

This approach to metadata also acknowledges that Western institutions do not have the knowledge, authority, or proficiency to catalog their items with Indigenous metadata schemata.

---

[29] Maureen Whalen, 2016, "Rights Metadata Made Simple" In *Introduction to Metadata*, by Murtha Baca, http://www.getty.edu/publications/intrometadata.
[30] Christen, "Opening Archives: Respectful Repatriation," 208-9.
[31] "Digital Heritage Metadata Fields," Mukurtu CMS, http://support.mukurtu.org/customer/en/portal/articles/2558813-digital-heritage-metadata-fields?b_id=633. (Accessed May 31, 2019).
[32] Christen, "Opening Archives: Respectful Repatriation," 201.
[33] Ibid.

Unless they hired catalogers and registrars from each community represented in their collection, the Fowler would not be able or skilled enough to apply Indigenous metadata schemata to their collections. Integrating their Western, Dublin Core metadata records alongside records created by communities "maintains the integrity of both institutional metadata and tribal community metadata while simultaneously showing the sharing of knowledge in multiple directions." [34]

Should the Fowler choose to use their Dublin Core records to repatriate items and collaborate with Indigenous communities through a platform like Mukurtu, there could be a reciprocal positive effect on both record sets. That is, incomplete Fowler records could be greatly improved by being placed alongside community records, which have the "local expertise, interpretation, and recollection" of communities. [35] In this way, digital repatriation, collaboration, and respectful community engagement benefit both institutions and Indigenous communities.

Further, Dublin Core's field for "rights" serves as administrative metadata that could document repatriation, a necessary logistical step in facilitating and building these collaborations.[36] This field could also be used to stipulate any specific access right restrictions. Some Indigenous communities understand private and public access differently than Western notions, as access is based "on a dynamic system of accountability where one's age, gender, ritual status, family, and place-based relationships all combine (and recombine as affiliations shift over a lifetime) to produce a continuum of access to materials within the community."[37] Should the Fowler wish to honor the Indigenous cultural practices of these items, they could document and implement access restrictions through the "rights" metadata field. Conceivably the record, then, would be accompanied by a blank photo, available for viewing on clearance.

In addition to applying access restrictions to their own digital collections, these restrictions would also apply to any sharing with digital repositories like Artstor and the Online Archive of California. Consistent and accurate use of this "rights" metadata field would thus be necessary for ensuring the responsible and respectful stewardship of these materials.

Adopting a standardized and highly interoperable metadata schema like Dublin Core also sets the stage for a post-custodial model, should the Fowler choose that mode of stewardship.

---

[34] Kimberly Christen, Alex Merrill, and Michael Wynne, 2017, "A Community of Relations: Mukurtu Hubs and Spokes," *D-Lib Magazine* 23 (5/6), https://doi.org/10.1045/may2017-christen.
[35] Peter Toner, 2004, "History, Memory and Music: The Repatriation of Digital Audio to Yolngu Communities, or, Memory as Metadata," Open Conference Systems: Sydney, Australia, 15.
[36] Gilliland, "Setting the Stage" In *Introduction to Metadata*.
[37] Christen, "Opening Archives: Respectful Repatriation," 189.

Through robust metadata records for objects, the Fowler could facilitate research and learning while still allowing for the physical objects to remain with their communities of origin. While a post-custodial model may be unlikely given the museum's financial stake in their holdings, establishing more thorough, consistent metadata practices at least allows for this option.

Proposed Controlled Vocabularies

There are several controlled vocabularies that the Fowler could utilize to standardize their metadata records, to ultimately make them more retrievable for users and interoperable with external repositories. The Getty Vocabularies, in particular, offer strong support for describing art, architecture, locations, artists, and museums. The Getty Vocabularies are multilingual and represent nearly 40 years of development and investment. Their broad scope and depth make them sustainable, reliable options for use for the Fowler's collections. For the Fowler, the Art & Architecture Thesaurus (AAT) could be used for the "format" field, when describing materials used, as well as in the "description" field, when describing the culture and style of a piece. The Getty Thesaurus of Geographic Names (TGN) could be used in the "coverage" field, to describe the spatial location of the object. The Union List of Artist Names (ULAN) could be used in the "publisher" field to describe the Fowler, as well as when applicable in the "creator" field.

In addition to the Getty Vocabularies, the DCMI Type Vocabulary should be used in the "type" field to ensure optimum integration and compatibility for searching aggregated records.[38] The "language" field, too, should use the recommended standards for Dublin Core: RFC 3066 and ISO 39, which define primary language tags and subtags.[39] For the "date" field, while the broader date ranges for most of the Fowler's materials do not lend themselves for easy adoption of Dublin Core's recommended SO 8601 standard, it is possible to use AAT's specification between types of date (alternative, inclusive, and coverage) where applicable.[40] At the very least, dates should be formatted consistently, even when a controlled vocabulary is not applicable.

Controlled vocabularies would primarily benefit the researcher user base. Resources like Artstor, the Google Cultural Institute, and the Online Archive of California aggregate digital collections from different museums in order to facilitate cross-repository searching. Ideally, this

---

[38] "Using Dublin Core," Dublin Core Metadata Initiative, http://www.dublincore.org/specifications/dublin-core/usageguide/elements/. (Accessed June 3, 2019).
[39] "Using Dublin Core," Dublin Core Metadata Initiative.
[40] Ibid.

would mean that researchers could more efficiently and more broadly locate relevant materials. In practice, however, keyword searching in these large repositories can be "woefully inadequate" due to the varied metadata practices of contributing institutions.[41] Utilizing the Getty Vocabularies within the Fowler's metadata records would improve the retrieval of their items within these repositories; Artstor, for example, uses ULAN and TGN.[42]

There are a few drawbacks to the Getty Vocabularies to consider with regard to the Fowler's collections. Some museum professionals have found AAT to have a steep learning curve, particularly when navigating its hierarchical structure to identify relevant terms.[43] Especially given the Fowler's financial and staffing constraints, this could present a real issue. Accordingly, this metadata strategy proposal limits AAT's required usage to "format," as materials terms are more straightforward than topical and conceptual terms. While this proposal suggests integrating topical and conceptual AAT terms alongside the label text in "description," this is optional. Additionally, this proposal is not requiring the "subject" Dublin Core field, which would utilize these more complex AAT terms, given the Fowler's financial and staffing constraints.

Further, the Getty Vocabularies can be biased toward Western art and architecture—a problem for the Fowler, as the majority of their collection is non-Western. This particularly can be felt with ULAN.[44] Similarly, these thesauri privilege art and architectural concepts, and may be less relevant to some of the Fowler's users, such as anthropologists, who would understand the Fowler's collections within a different framework. While these biases are not ideal, the Getty Vocabularies are the most robust thesauri available for collections of cultural artifacts. Some of the bias of ULAN can be circumvented, as many of the artists within the Fowler's collections are unknown. Ultimately, the potential benefits with regard to aggregated search, retrievability, and accessibility make using the Getty Vocabularies a fruitful strategy.

Finally, on a more practical level, in order to encourage users of the Fowler's digital collections to enjoy the full benefits of the controlled vocabulary, it is advised that the Fowler includes a section on their website linking to the Getty Vocabularies and explaining how they

---

[41] Murtha Baca and Melissa Gill, 2015, "Encoding Multilingual Knowledge Systems in the Digital Age: The Getty Vocabularies," *Knowledge Organization* 42 (4), 232.
[42] "Metadata Policy," Artstor.
[43] Alison Gilchrest, 2003, "Factors Affecting Controlled Vocabulary Usage in Art Museum Information Systems," *Art Documentation: Journal of the Art Libraries Society of North America* 22 (1), 15.
[44] Ibid.

were applied to the digital collections. This would empower users to create more effective and targeted searches.

Sample Record

The following figure shows a transformation of a current record[45] at the Fowler, utilizing the proposed metadata schema and controlled vocabulary described above.

CURRENT RECORD	
Object Name:	Lambayeque vessel
Artist:	Unknown
Cultural Group:	Lambayeque
Place of Origin:	Peru, north coast
Date:	900 – 1300 C.E.
Materials Used:	Ceramic
Dimensions:	H: 13.50 cm, L: 14.50 cm, W: 12.00 cm
Credit line and Accession Number:	Fowler Museum at UCLA. Gift of Dr. Harry and Claire Steinberg. X92.311
TAG:	Andean Ceramics

TRANSFORMED RECORD	
Title:	Lambayeque vessel
Creator:	unknown Lambayeque°
Date:	900 – 1300 C.E.
Format:	Ceramic (material)* H: 13.50 cm, L: 14.50 cm, W: 12.00 cm
Description:	[Label text; whether or not item is on view]. Originated from the Lambayeque peoples* of the north coast of Peru.
Coverage:	Lambayeque Peru
Identifier:	Accession number X92.311
Publisher:	Fowler Museum of Cultural History°
Type:	PhysicalObject¬
Rights:	Credit line: Fowler Museum at UCLA. Gift of Dr. Harry and Claire Steinberg. [Access rights; repatriation records as applicable].
TAG:	Andean Ceramics

Vocabulary Key:
° = ULAN
* = AAT
 = TGN
¬ = DCMI Type

As shown in the above figure, the Fowler's current metadata makes a fairly easy transition to a Dublin Core record. One of the more detailed object records was chosen to show the full potential of the transformed record. For records with less detail—such as those missing dates—some research will be required, as indicated in this record by the brackets. If the Fowler prefers the front-end to use different terms than Dublin Core's (such as preferring "place of

[45] "X92.311 Lambayeque Vessel," Fowler Museum.

origin" to "coverage"), they can always choose to encode the Dublin Core schema into the back-end, and have the front-end display the preferred term.

While most fields mapped on clearly, a few are a bit more involved. The "description" field, for example, has been inconsistent across the current records, with some including curatorial label text and others specifying within the record more generally if it is currently on view in the museum. In addition to these two elements, this new strategy proposes integrating AAT controlled terms where possible to increase searchability, especially as the field represents a "potentially rich source of indexable terms."[46] Since the "description" field utilizes full sentences to present more in-depth information to the user, elements that are lost through the implementation of controlled vocabularies in other fields can be incorporated here. For example, the place of origin was initially described as "Peru, northern coast;" however, "coverage" does not include the coastal detail as "Peru, northern coast" was not the preferred term for the region in TGN. This information can still be saved and presented to the user, under "description."

The "rights" field is a particularly valuable addition to the collection record. In addition to the credit line, this would be a useful field for supporting repatriation efforts, detailing the object's repatriation history as well as any information as to whether the source can be viewed or accessed by the general public. On the research side, once the Fowler determines the rights status for their items, they could evaluate the premade, standardized rights statements provided by Rightsstatements.org, and determine which is applicable to the object. Then, within the "rights" field, the Fowler could link directly to the Righststatements.org statement. Developed in part by the Digital Public Library of America and Europeana, these rights statements use clear, standardized language to promote engagement with materials and repository aggregation.[47]

The fields in the above transformed record are all required as part of the proposed metadata strategy, aside from the "TAG" field. Tagging within the collection is currently limited and decentralized, and thus not exceptionally usable as a mode of information discovery. It appears that only a select number of items were tagged—the Andean ceramics, Andean textiles, and "Fowler at Fifty." While it would be useful if all of the items online could be tagged to facilitate discovery and search, that would require significant planning and research on user

[46] "Using Dublin Core," Dublin Core Metadata Initiative.
[47] International Rights Statements Working Group, 2017. "Recommendations for Standardized International Rights Statements," White Paper Version 1.2, 4.

search patterns. Thus, a simpler approach to searching and browsing would be to rely instead on the collection filters on the left navigation bar of the digital collections, which currently include geographic region, culture, date/era, and medium—all of which are already detailed in the metadata. These filters are preferable to the tags, which currently are not consistent nor comprehensive enough to provide valuable and thoughtful points for discovery and access.

The "relation" and "source" fields from Dublin Core are also optional, as they are largely not applicable to most pieces, and instead reflect Dublin Core's strength in describing digitized bibliographic materials. The "contributor" field is optional as well, seeing as it is not applicable to most of the Fowler's holdings. As described above, the "subject" field is optional as it would require more time and money; for a limited approach, the "description" field will offer enough context to make the resources usable. The "language" field is also optional, as it is not relevant to many of the materials. While some do have writing on them, others do not, with some coming from cultures that did not have a written language.

Metadata Creation Logistics

Implementing this metadata strategy will require the work of information or collections professionals. While Dublin Core is the most basic schema for digital materials, it still may feel foreign to the layperson, especially when integrating various vocabularies with it. Additionally, as research will be required to complete some of these records, either within internal systems or externally, it would be best for a museum or information professional to do this work. That said, this work does not require a senior professional, and could be completed by a more junior professional within the field. Perhaps given the Fowler's physical location on campus, this would be an apt opportunity for graduate students in Information Studies. Hiring student workers would also be an economical option for the Fowler, while providing students with valuable experience.

It also may be worth looking into the functionality of the Fowler's collections management system, to discover how much of this metadata could be automatically extracted from this system. Automating metadata creation by exporting metadata from the Fowler's collections management system as a CSV file, for example, could save significant time and money. If their collections management system allows for this, the metadata the Fowler has already created within that system could be exported, cleaned and standardized within OpenRefine, and then imported into the digital collections' content management system as

structured metadata fields. While cleanup would still be involved, this would prevent duplicating some of the metadata entry already done on one of the Fowler's systems.

A potential drawback to using graduate students for description would be the high turnover that necessarily comes with a two-year master's program. This could result in inconsistencies, such as uneven application of vocabularies or simple errors that come during training periods, which would happen more frequently as new people cycle in and out. These problems with inconsistency would be exacerbated by the fact that the museum currently does not have anyone on staff with an MLIS degree. Strong training documentation would be necessary, then, complete with workflow diagrams, example records, and robust resources and reference materials. Additionally, internal records could be maintained as to which vocabulary terms the Fowler is using within the Getty Vocabularies. Some content management systems like Drupal even allow for integration of vocabularies within the system to facilitate this tracking and encourage consistent implementation.[48] All training and reference materials should be constantly updated and improved upon, as any institutional knowledge will likely be temporary.

Conclusion

Digital collections for museums represent a real opportunity to further access, discovery, and engagement with materials. For the Fowler, this opportunity is magnified due to the nature of their collections. On the research side, not just art historians are interested in the Fowler's materials—scholars from anthropology, archaeology, ethnography, and architecture could all utilize the Fowler's collections to further their research. On the public side, Indigenous and marginalized communities could be reunited with items of cultural heritage through repatriation, as the Fowler's holdings include materials from other countries, including spiritual and cultural materials from Indigenous communities. Metadata could play a key role in serving both of these user bases. While the objectives of these two user bases may seem at odds with one another—with researchers needing more access and aggregation, and repatriation at times involving access restrictions—metadata can play a critical part in answering the needs of both communities, simultaneously facilitating more meaningful search and access of the collection at large, while also ensuring the safety and cultural repatriation of select materials.

[48] "Web Taxonomy Plugin for Getty Vocabularies," 2014, Drupal.Org, https://www.drupal.org/project/wt_getty.

References

"Archaeology." n.d. Fowler Museum. Accessed June 1, 2019. https://www.fowler.ucla.edu/archaeology/.

Baca, Murtha, and Melissa Gill. 2015. "Encoding Multilingual Knowledge Systems in the Digital Age: The Getty Vocabularies." *Knowledge Organization* 42 (4): 232–43.

Baca, Murtha, Patricia Harpring, Elisa Lanzi, Linda McRae, and Ann Whiteside, eds. 2006. "Cataloging Cultural Objects: A Guide to Describing Cultural Works and Their Images." American Library Association. http://cco.vrafoundation.org/index.php/toolkit/cco_pdf_version/.

"Categories for the Description of Works of Art (CDWA)." n.d. Getty Research Institute. Accessed June 2, 2019. http://www.getty.edu/research/publications/electronic_publications/cdwa/introduction.html#general.

Christen, Kimberly. 2011. "Opening Archives: Respectful Repatriation." *The American Archivist* 74 (Spring/Summer 2011): 185–210.

Christen, Kimberly, Alex Merrill, and Michael Wynne. 2017. "A Community of Relations: Mukurtu Hubs and Spokes." *D-Lib Magazine* 23 (5/6). https://doi.org/10.1045/may2017-christen.

"Collections Overview." n.d. Fowler Museum. Accessed May 31, 2019. https://www.fowler.ucla.edu/collections/home/.

"Digital Heritage Metadata Fields." n.d. Mukurtu CMS. Accessed May 31, 2019. http://support.mukurtu.org/customer/en/portal/articles/2558813-digital-heritage-metadata-fields?b_id=633.

"Document Search Results for '"Notice of Intent To Repatriate Cultural Items" Fowler.'" n.d. Federal Register. Accessed June 2, 2019. https://www.federalregister.gov/documents/search?conditions%5Bterm%5D=%22Notice+of+Intent+To+Repatriate+Cultural+Items%22+fowler.

Gilchrest, Alison. 2003. "Factors Affecting Controlled Vocabulary Usage in Art Museum Information Systems." *Art Documentation: Journal of the Art Libraries Society of North America* 22 (1): 13–20.

Gilliland, Anne. 2016. “Setting the Stage.” In *Introduction to Metadata*, by Murtha Baca. http://www.getty.edu/publications/intrometadata.

International Rights Statements Working Group. 2017. “Recommendations for Standardized International Rights Statements.” White Paper Version 1.2.

“Lega Figures.” n.d. Fowler Museum. Accessed May 31, 2019. https://www.fowler.ucla.edu/collections/lega-figures/.

“Metadata Policy.” n.d. Artstor. Accessed June 4, 2019. https://www.artstor.org/contribute/metadata-policy/.

Miller, Stephen J. 2011. *Metadata for Digital Collections: A How-to-Do-It Manual*. London, United Kingdom: Facet Publishing.

Muchnic, Suzanne. 2013. “UCLA's Fowler Museum Turns 50 in Worldly Fashion.” *Los Angeles Times*. Los Angeles Times. September 28. https://www.latimes.com/entertainment/arts/la-xpm-2013-sep-28-la-et-cm-ucla-fowler-museum-50-20130929-story.html.

“Open Graph and Its Impact on SEO.” n.d. Yakaferci. Accessed May 31, 2019. //www.yakaferci.io/open-graph/.

“Open Graph Protocol.” n.d. Accessed May 31, 2019. http://ogp.me/.

“Products Archive.” n.d. Fowler Museum. Accessed May 31, 2019. https://www.fowler.ucla.edu/collections/.

“Responsive Web Design and Development Services.” n.d. Citrus Studios. Accessed June 1, 2019. https://www.citrusstudios.com/online-marketing-services/responsive-web-design-development/.

Riley, Jenn. 2010. “Glossary of Metadata Standards.” http://jennriley.com/metadatamap/seeingstandards_glossary_pamphlet.pdf.

“Technical Overview.” n.d. Online Archive of California (OAC) / Calisphere Contributor Help Center. Accessed June 4, 2019. https://help.oac.cdlib.org/support/solutions/articles/9000081989-technical-overview.

Toner, Peter. 2004. “History, Memory and Music: The Repatriation of Digital Audio to Yolngu Communities, or, Memory as Metadata.” Open Conference Systems: Sydney, Australia. https://ses.library.usyd.edu.au/handle/2123/1518.

"Using Dublin Core." n.d. Dublin Core Metadata Initiative. Accessed June 3, 2019. http://www.dublincore.org/specifications/dublin-core/usageguide/elements/.

"VRA Core 4.0 Introduction." 2014. http://www.loc.gov/standards/vracore/VRA_Core4_Intro.pdf.

"Web Taxonomy Plugin for Getty Vocabularies." 2014. Drupal.Org. August 27, 2014. https://www.drupal.org/project/wt_getty.

Whalen, Maureen. 2016. "Rights Metadata Made Simple." In *Introduction to Metadata*, by Murtha Baca. http://www.getty.edu/publications/intrometadata.

"World Arts and Cultures Courses." n.d. UCLA General Catalog 2018-19. Accessed June 3, 2019. https://catalog.registrar.ucla.edu/ucla-catalog18-19-1398.html.

"X91.410 Drinking Horn." n.d. Fowler Museum. Accessed May 31, 2019. https://www.fowler.ucla.edu/product/x91-410-drinking-horn/.

"X95.38.207a,b Betel Mortar." n.d. Fowler Museum. Accessed May 31, 2019. https://www.fowler.ucla.edu/product/x95-38-207ab-betel-mortar/.

"X92.311 Lambayeque Vessel." n.d. Fowler Museum. Accessed May 31, 2019. https://www.fowler.ucla.edu/product/x92-311-lambayeque-vessel/.

"X2007.21.90 Lega Spoon." n.d. Fowler Museum. Accessed May 31, 2019. https://www.fowler.ucla.edu/product/x2007-21-90-lega-spoon/.

Elective Coursework

Digital Humanities // DH 201
Miriam Posner
Fall 2019

Abstract
The Carnegie Museum of Art (CMOA) is a contemporary art museum in Pittsburgh, Pennsylvania. CMOA asserts that their programming, exhibits, and publications “frequently explore the role of art and artists in confronting key social issues of our time, combining and juxtaposing local and global perspectives.” Their mission statement champions creativity with a global focus, stating that “CMOA collects, preserves, and presents artworks from around the world to inspire, sustain, and provoke discussion, and to engage and reflect multiple audiences.”

This project takes its lead from CMOA's mission statement, examining their accession records to get a sense of how global their collection is. CMOA recorded the nationality of artists within each artwork’s accession record, and was fairly thorough in this effort. While there is no silver bullet for determining the diversity of a museum’s collection, the data visualizations in this project will tease apart aspects of CMOA’s contemporary art collection to get a better understanding of the nationalities and countries represented in their artworks.

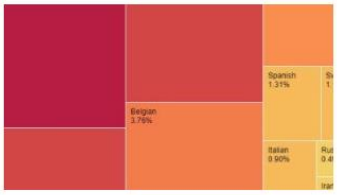
Inclusion in portfolio
This project required strong data cleaning and visualization skills, and developed my expertise in OpenRefine, Tableau, and web design. Further, a recurrent theme of the course was debunking the myth of "raw data." Data need to be collected, cleaned, and curated, all of which involve human intervention. A large part of this project was thus documenting decisions I made when working with data in order to allow people to critically engage with and analyze my work. I included this project in my portfolio as it reflects my ability to work with data in a transparent, ethical manner, and also illustrates my technical skills with data work.

A printed PDF version is included here. To see the website with its interactive data visualizations, see savannahlake.github.io/dh201.

CREATIVITY + CONTEXT

The Carnegie Museum of Art aims to present "artworks from around the world to inspire, sustain, and provoke discussion, and to engage and reflect multiple audiences." This project investigates one aspect of this effort, exploring international representation within the museum's contemporary art collection.

ABOUT THE PROJECT | DATA VISUALIZATIONS

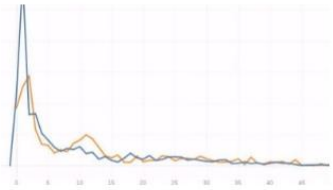


Mapping Artists

What are the artists' nationalities?

Explore the nationalities of the artists and artworks through interactive maps and graphs that depict the percentage breakdown of the collection by nationality.

MORE



Acquistion Trends

What is purchased and what is donated?

Consider museum priorities through graphs that depict purchasing trends, as well as visualizations that show the time between an artwork's creation and its acquisition by the museum.

MORE

Items on View

Who is on view at the museum?

Perhaps the most important measure is whether or not an artwork is locked away in storage or on view at the museum for visitors to see. Learn what artworks are on view by nationality.

MORE

The Ambiguity of Nationality

What can't be captured in the data?

What are the limitations of examining an art collection by artist nationality? Tease apart what "nationality" means, and assess how much the data is able to capture.

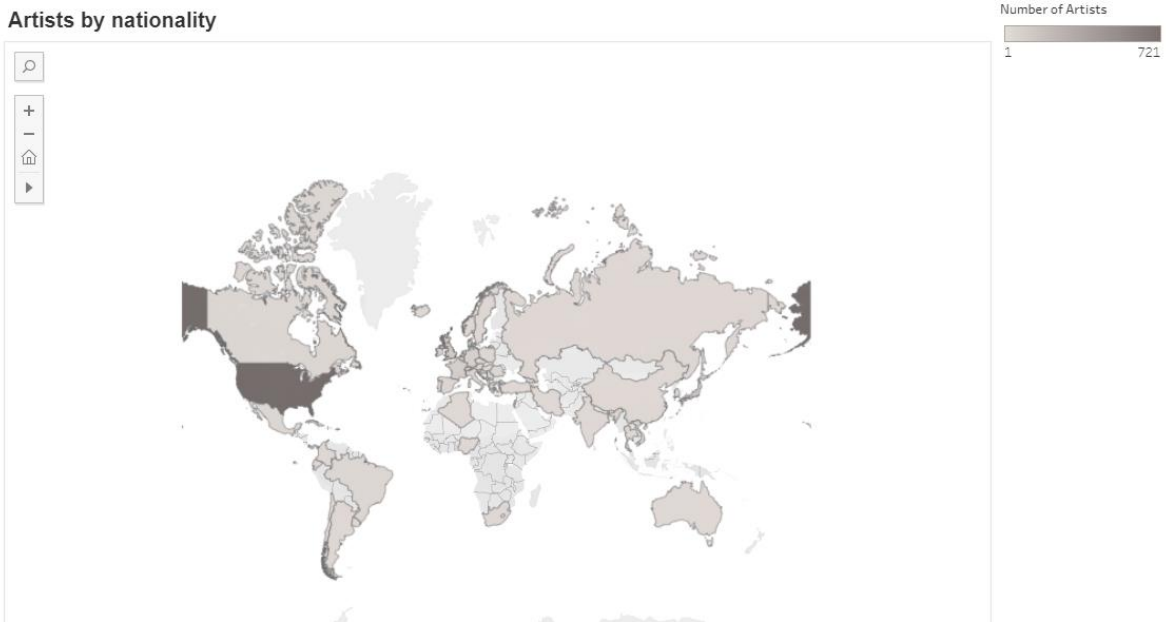
MORE

MAPPING ARTISTS

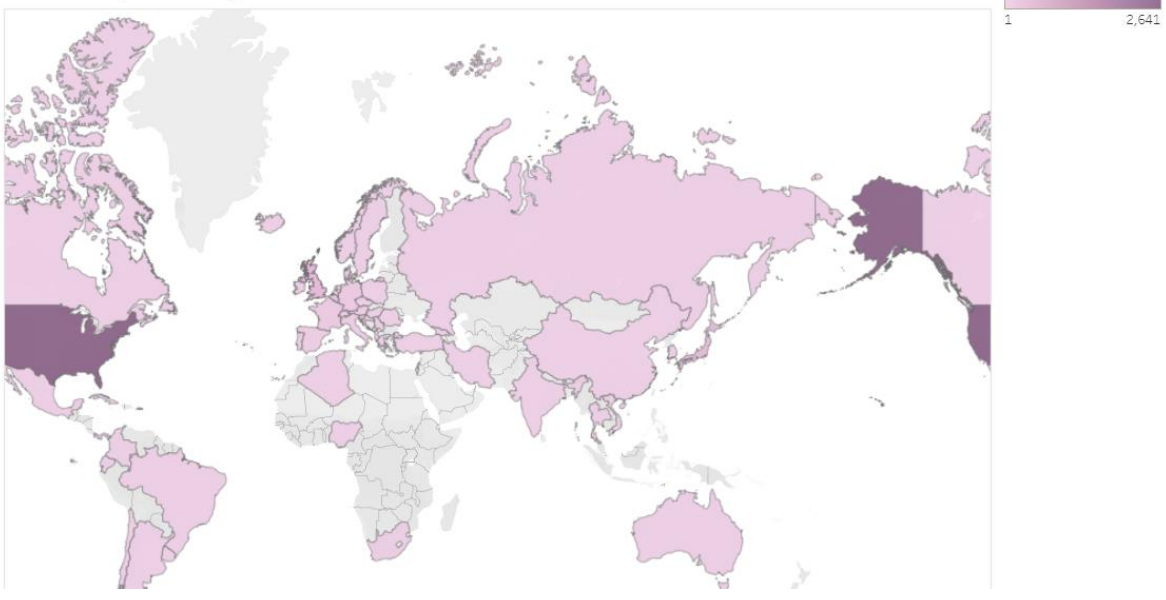
The contemporary art department represents 1,304 artists and 4,869 artworks. Where are these artists and artworks from? Below are maps of both the artists and the artworks by nationality. While the maps are similar, there are some striking distinctions. Interacting with the map by hovering over the country will show you the number or artists and artworks associated with that country. You'll notice that the correlation between the number of artworks and the number of artists varies by country. For example, the collection has one artist from Algeria and one artist from Serbia, but the collection features 15 pieces from the Serbian artist and only one from the Algerian artist. Similarly, the collection has two artists from Iran, Colombia, and Vietnam, but the number of artworks varies, with 23 artworks coming from the Iranian artists, 7 artworks from the Colombian artists, and 2 artworks from the Vietnamese artists.

The issue of representation, then, is nuanced by volume of artworks versus artists. Could 15 artworks from one Serbian artist represent a commitment to that one artist as opposed to Serbian art in general? Is having 13 artists contribute one piece each any different or better, as seen in the case of Canada?

Artists by nationality

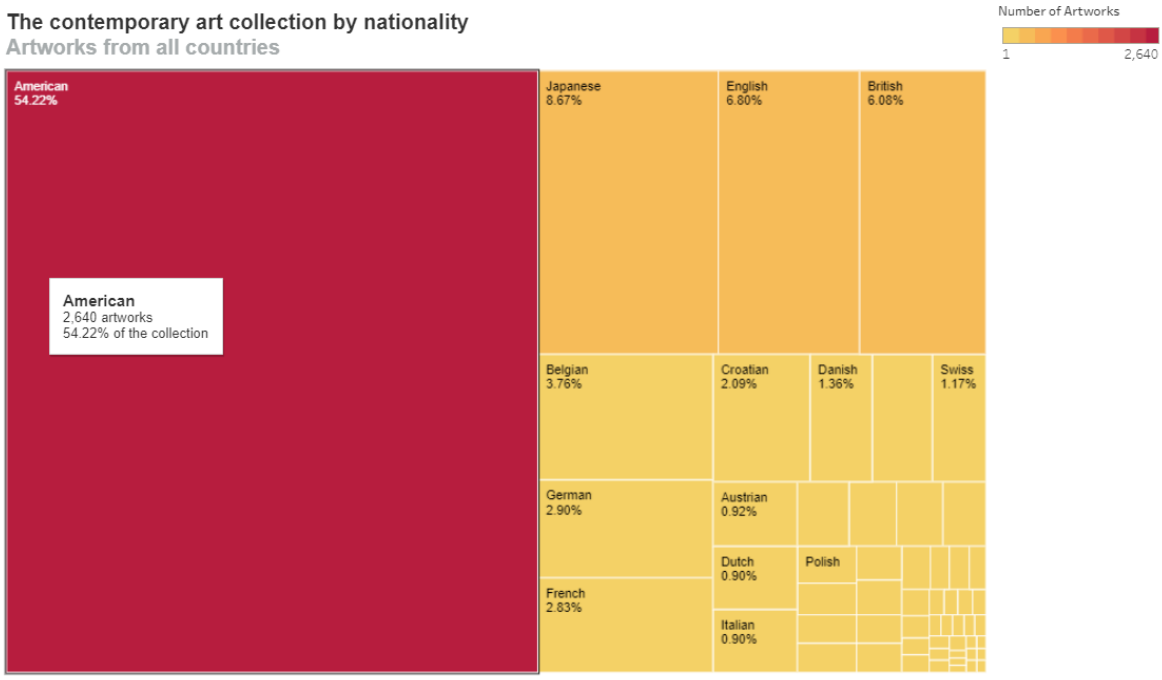


Artworks by nationality



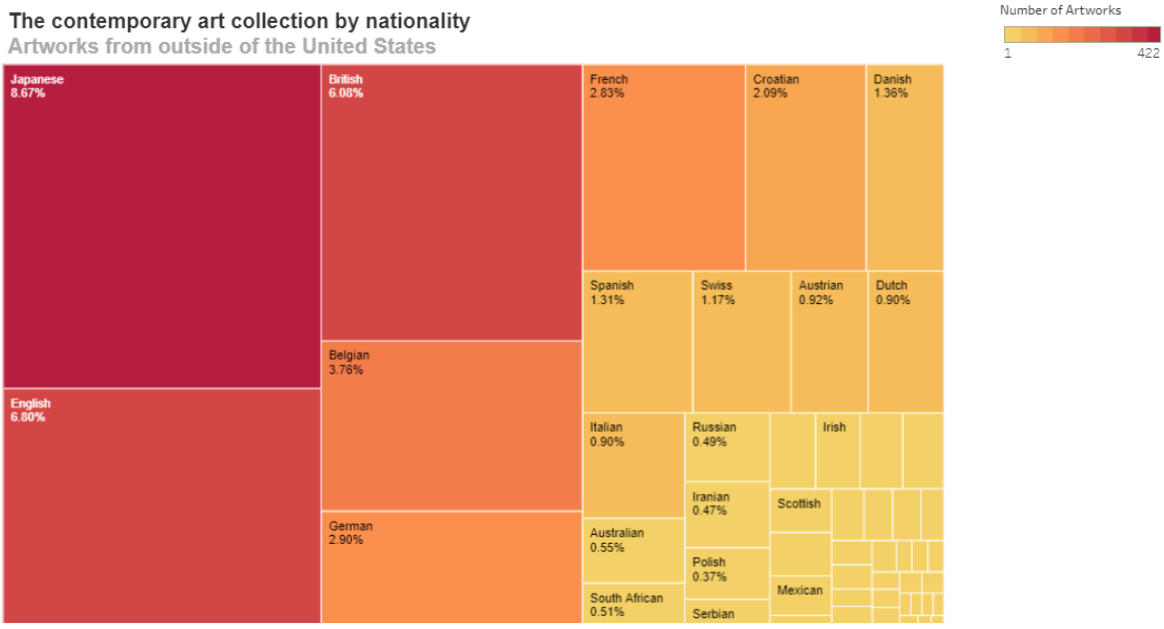
Click here for details on how these visualizations were created.

Overwhelmingly though, the maps indicate that the majority of the artists and artworks are coming from the United States, with 721 artists and 2,640 artworks. That represents 55.3% of the artists and 54.2% of the artworks. While the volume of artwork from the United States may seem high compared to any other particular country, nearly half the collection is from outside of the United States, suggesting a concerted effort to collect from around the world more generally.



Click here for details on how these visualizations were created.

Suppressing the United States from the data is able to give a clearer picture of the distribution of nationalities represented in the collection outside of the United States. CMOA's strongest holdings are in Japanese art, followed mostly by art by artists from western Europe. Australia is the first country to leave Europe after Japan, and is the 14th most collected nationality outside of the United States. Accordingly, while CMOA has stronger Japanese holdings (comprising 8.67% of their overall collection), when it comes to collecting artworks from nationalities from outside of the United States, they collect most actively in Europe, with their first 12 European countries alone comprising 31.02% of the overall collection.



Click here for details on how these visualizations were created.

ACQUISITION TRENDS

Acquiring an artwork often involves a financial commitment, and can reflect the priorities of the museum at large. Acquisitions also build the collection and play a part in what museum visitors get to see. I wanted to tease out two trends in the data with regard to acquisition and artist nationality: purchasing decisions and latency.

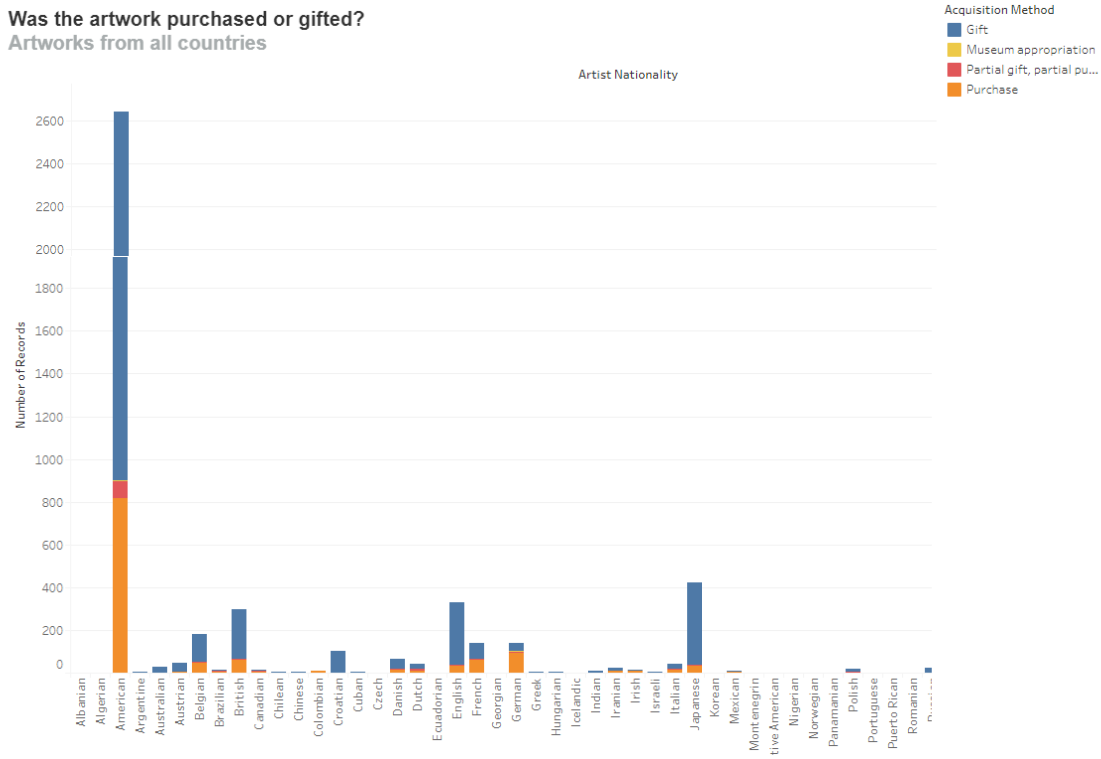
Purchasing decisions

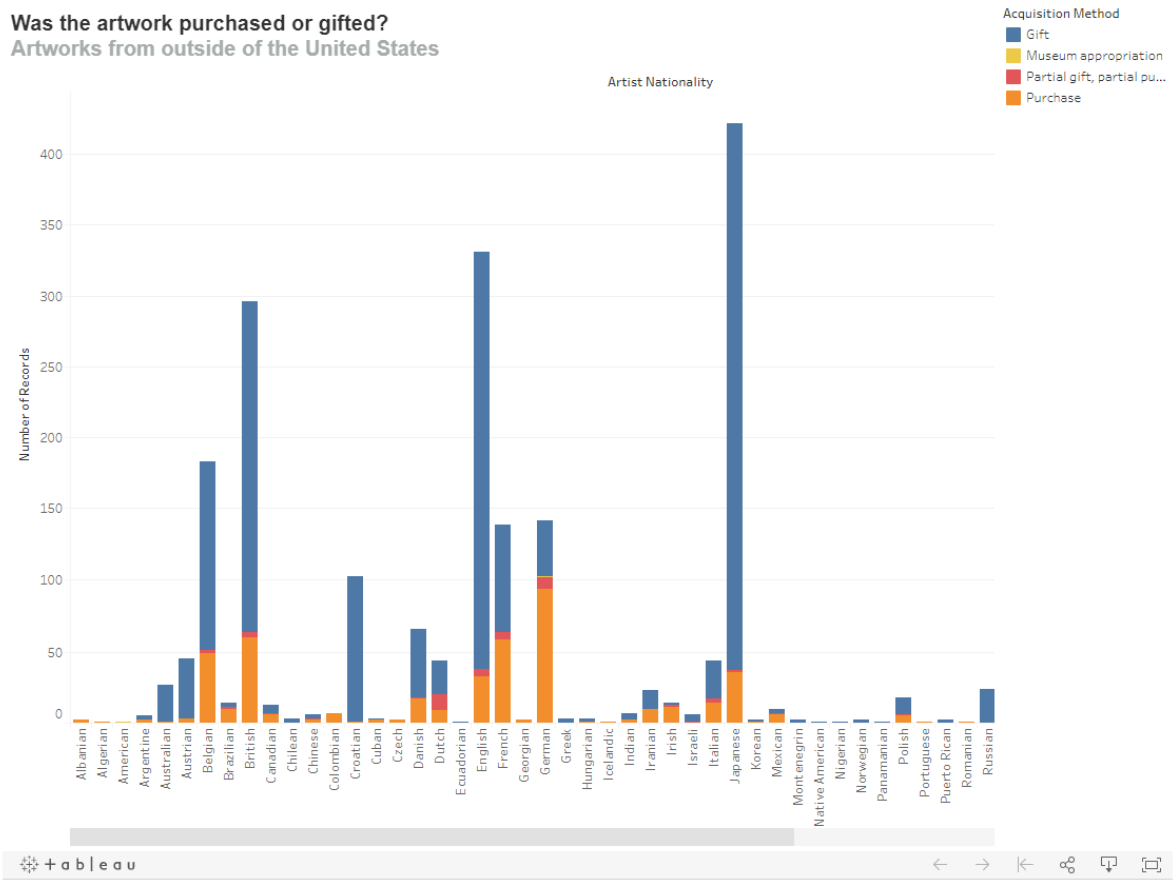
With regard to purchasing decisions, artworks can either be purchased by the museum, gifted to the museum, or acquired with funds both from the museum and a gift. The majority of CMOA's artworks were gifted, with roughly two-thirds of their holdings gifted and one-third purchased.

Click here for details on how this visualization was created.

Of its 4,689 artworks, the CMOA dataset classified two of the artworks as being acquired through "museum appropriation." Unfortunately, the CMOA dataset does not have a data dictionary defining what that means, but per the Getty Art and Architecture Thesaurus, these funds likely refer to the museum's budget.

This breakdown—of two-thirds gifted, one-third purchased—plays out largely in the data. The following bar charts break down artworks by acquisition method, with the second focused on countries outside of the United States to give greater visibility to their data. There are a few notable exceptions—German, South African, and Turkish art, for example, are almost exclusively purchased, not gifted.





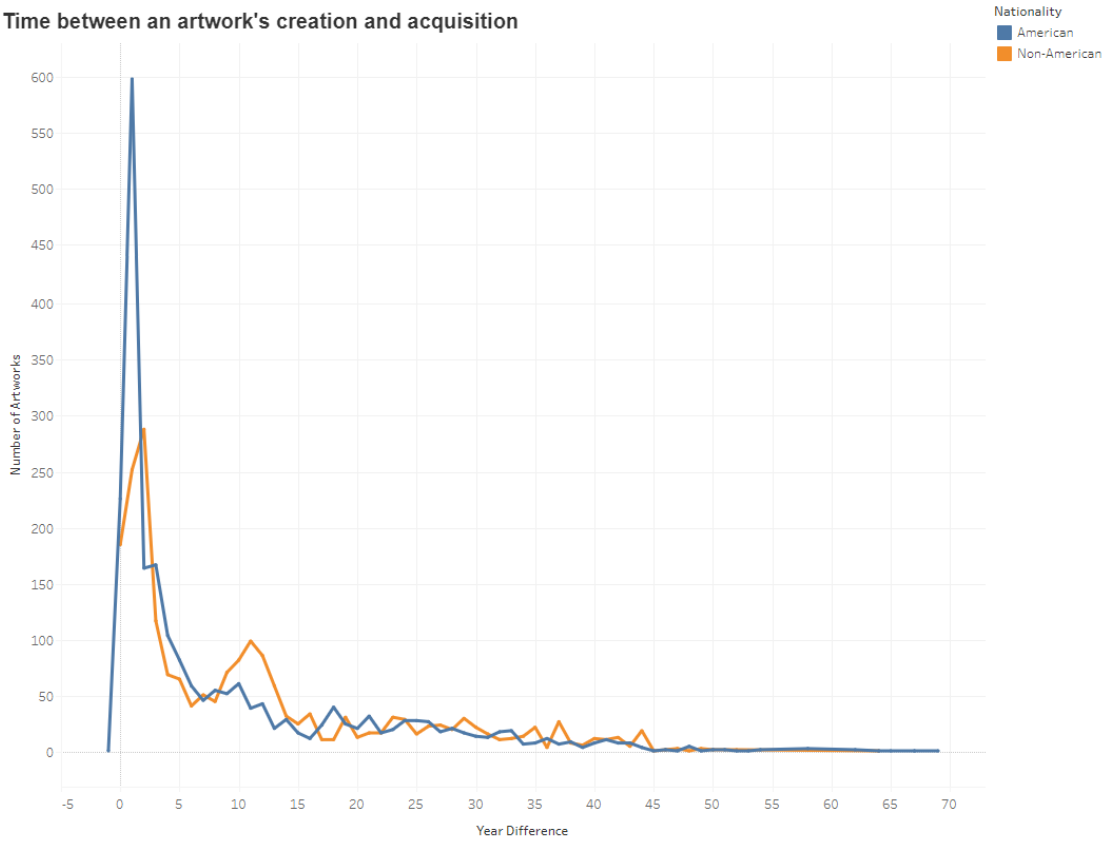
Click here for details on how these visualizations were created.

Artworks that have been purchased, as opposed to gifted, could reflect a more sincere investment in acquiring artwork from that region, as the museum is putting its own funds towards that purchase. From this viewpoint, the museum seems to be invested in acquiring German art, which has both a healthier volume and is weighted towards purchased items over gifted items. However, while only 30.9% of art from the United States was purchased, that represents 816 artworks—which is more than the total number of artworks, purchased or otherwise, from any other country. In order of volume, the top five most represented nationalities after American art are Japanese art with 422 artworks, then English with 331 artworks, then British with 296 artworks, and Belgian with 183 artworks—all of which are a fraction of the United States' 816 purchased artworks, and an even smaller fraction of the 2,640 total American artworks. Clearly, the largest financial investment in artwork by nationality is in American artwork.

Latency

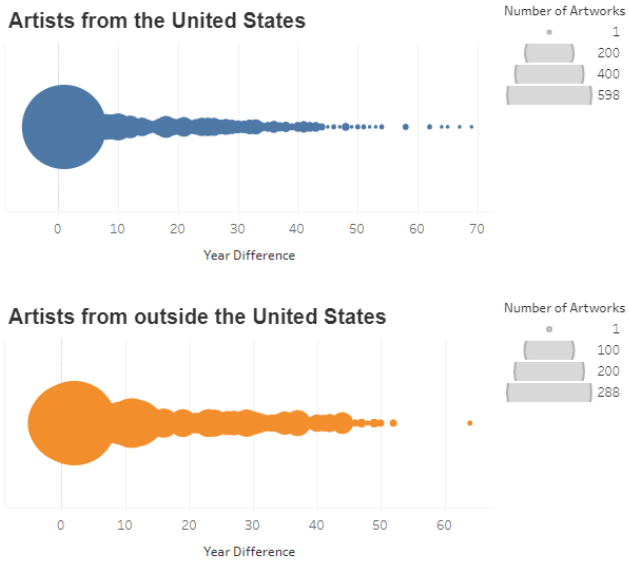
As another view into the museum's commitment to art from a particular region, I examined the length of time between an artwork's creation and its acquisition. That is, if artworks are generally bought shortly after their creation in a certain region, it could suggest that the museum is keeping tabs on that art scene and is actively searching for pieces and acquiring them. Conversely, if artworks from a region are generally purchased decades after their creation, that could suggest that the museum is newly aware or newly interested in this art scene. This is by no means a perfect measure, as there could be any number of reasons affecting an artwork's acquisition. However, general trends in the data may speak to acquisition priorities.

Both artworks from the United States and artworks from elsewhere follow similar acquisition trends, with most pieces being purchased within five years of their creation.



Click here *for details on how this visualization was created.*

However, when the scale is adjusted so that the increased volume of artworks from the United States is controlled for, it is apparent that artworks from the United States are purchased more closely to their creation date than artworks from outside the United States. More specifically, within the United States the majority of artworks are purchased within the first six years, with purchasing thinning out thereafter. For artworks from outside of the United States, this tapering off is less dramatic, reflecting longer periods of time between creation and acquisition.



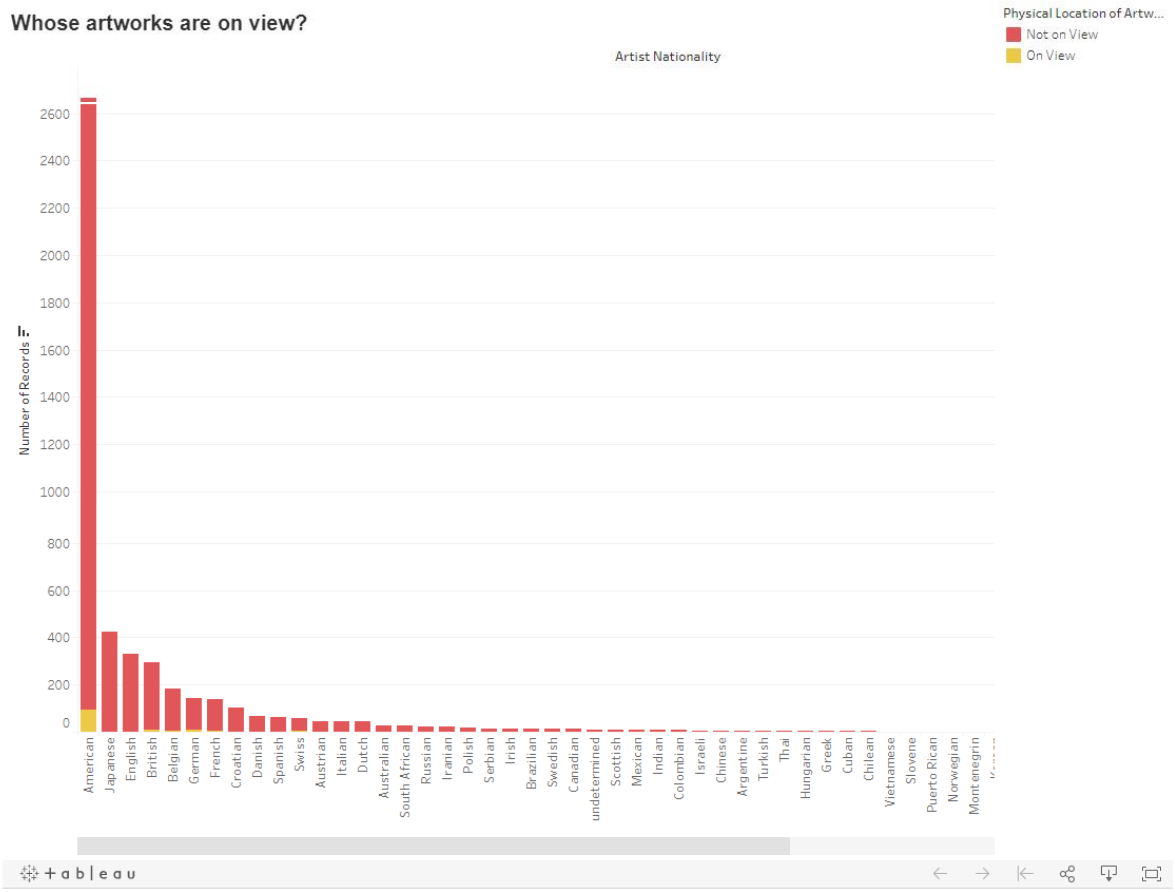
Click here *for details on how this visualization was created.*

The shorter time between an artwork's creation and acquisition date indicates more commitment and attention paid to American art scenes as opposed to those from other countries. However, the data shows that the latency periods—while different—are not dramatically dissimilar, suggesting that the museum is actively acquiring and following art scenes around the world.

ITEMS ON VIEW

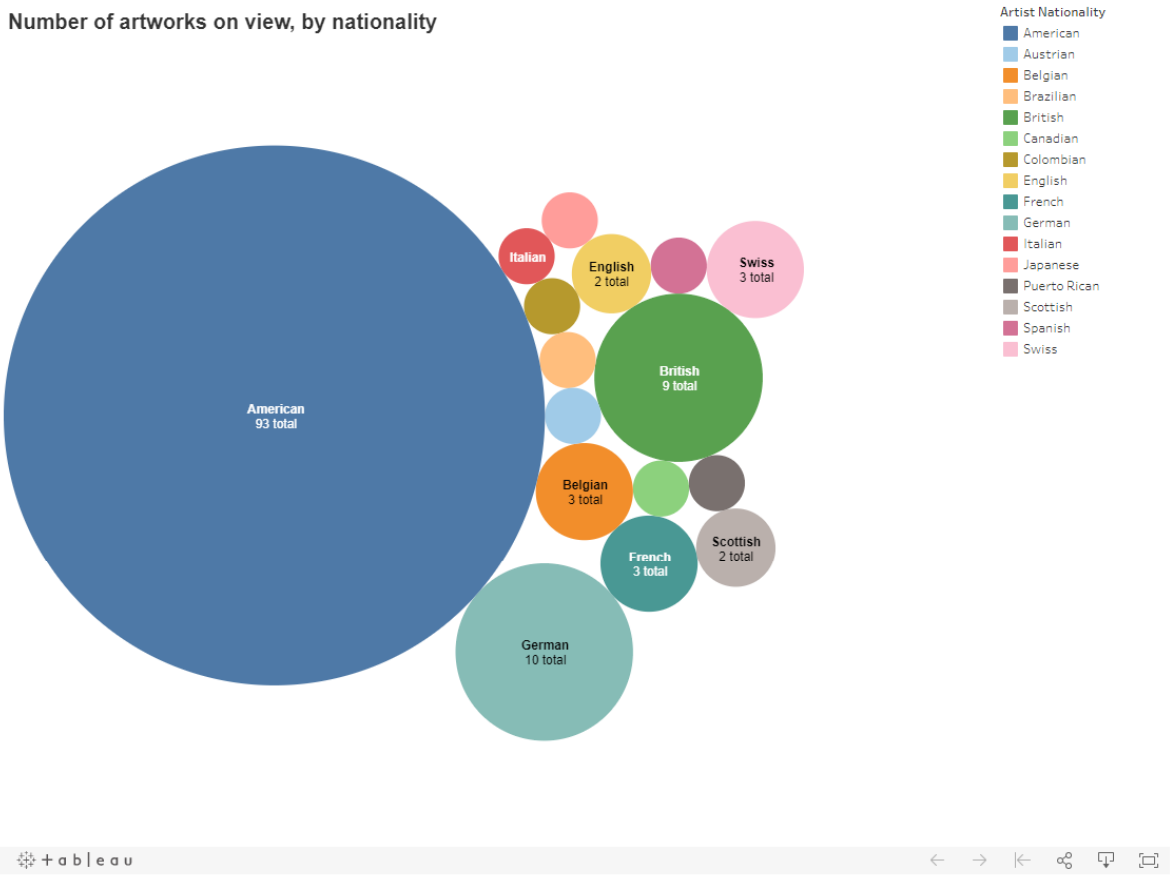
When trying to understand how global a collection is, it is not enough to look simply at what was acquired. CMOA's mission statement asserts that the museum "collects, preserves, and presents artworks from around the world," specifically to "inspire, sustain, and provoke discussion, and to engage and reflect multiple audiences." Their mission recognizes the importance of visitor engagement with art, which is only possible if the artwork is actually on view. It is important, then, to evaluate which pieces are on view, and in turn, which nationalities are represented.

Generally speaking, the vast majority of their contemporary art collection is not on view. Of their 4,869 artworks, only 133 are on view, or 2.7%.



Click here for details on how this visualization was created.

Of the 133 pieces that are on view, 93 of them are American, followed by 10 German artworks and 9 British artworks. Interestingly, despite their greater holdings in Japanese artworks (see *Mapping Artists* for a full breakdown of holdings by nationality), only one of these pieces is on view. Greater holdings thus does not always equate greater visibility. Overall, the representation is overwhelmingly American.

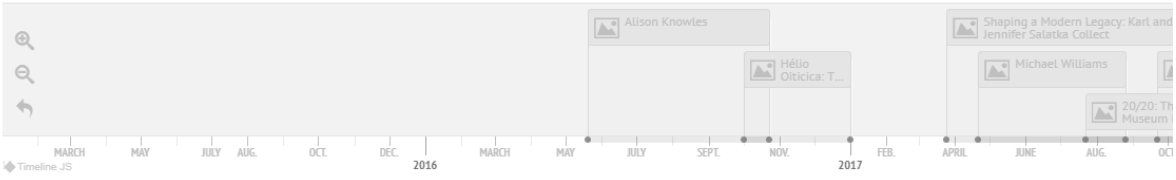


Click here for details on how this visualization was created.

In addition exploring their permanent collection, CMOA's exhibitions provide an opportunity to explore what art they choose to highlight, and what can be seen by the public. Of the twelve exhibitions held since 2016, nine of them featured exclusively American artists. The remaining three featured a Brazilian artist, German and American artists, and the final featured many artists from Austria, the Bahamas, Cameroon, Cherokee Nation, Colombia, England, Germany, Ghana, India, Japan, Jordan, Kenya, Korea, Kuwait, Lebanon, Navajo Nation, Nigeria, Nonuya Nation, Pakistan, Palestine, Scotland, Senegal, Switzerland, the United States, and Vietnam. While there was some international representation, generally speaking the exhibitions overwhelmingly feature American artists.

CONTEMPORARY ART EXHIBITIONS AT CMOA

A look into the nationalities represented in recent contemporary art exhibitions at CMOA, ranging from 2016 to now. For full descriptions of these exhibitions, see cmoa.org/exhibitions/

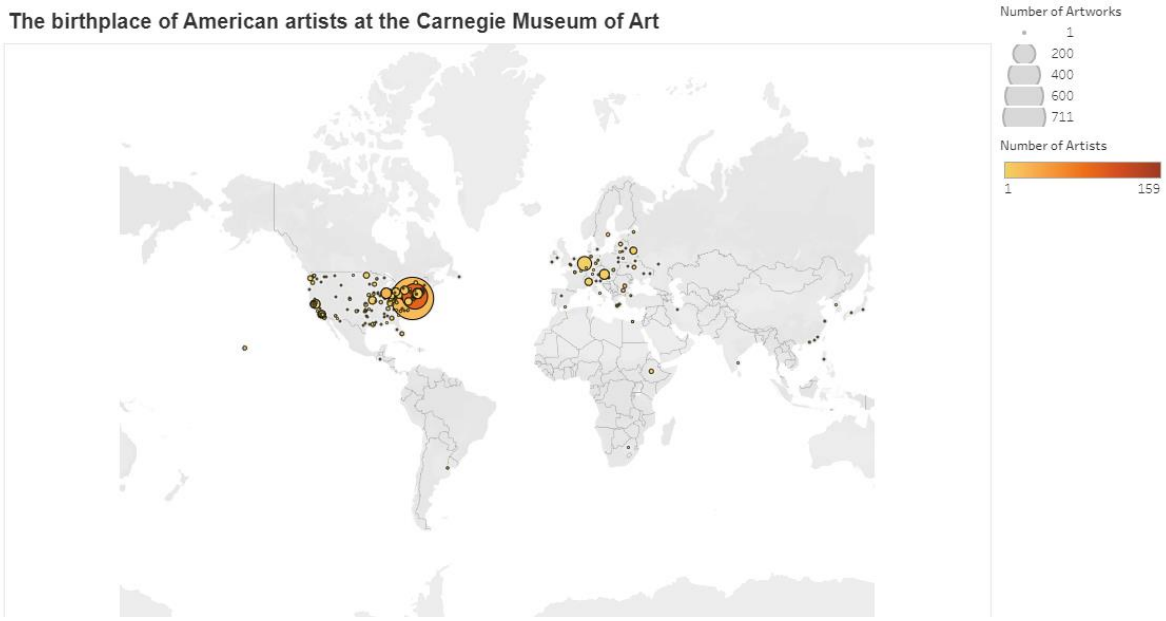


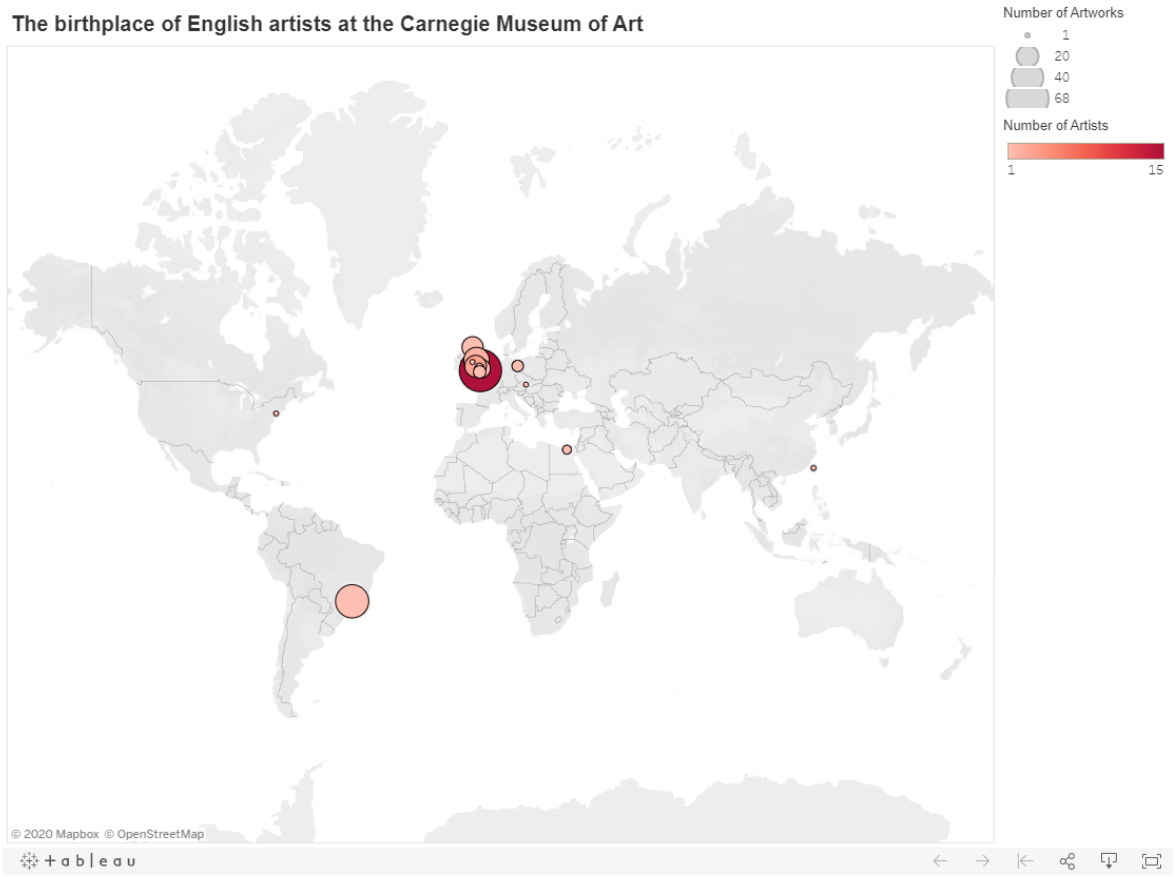
THE AMBIGUITY OF NATIONALITY

While exploring CMOA's collection by artist nationality offers an avenue into understanding how global the collection is, it is important to recognize the limitations of this data. Nationality does not equal diversity. The most common nationality within the collection—American—represents people of all different races, gender identities, abilities, and sexual preferences, which all shape their experience of the world. The fact that they are all the same nationality, American, does not mean that they all share the same lived experience. Nationality can only go so far in reflecting diversity of experience.

Further, there are people who may identify with multiple nationalities. Perhaps they lived in a different country and have immigrated elsewhere, or their family carries on traditions that makes them feel a connection with that country and identity. The data is unable to capture this complexity.

In an effort to tease out some of these complexities, I mapped the birthplace of artists classified with an American or an English nationality. While many of the artists were born in the United States and England, respectively, quite a few were born elsewhere—suggesting that perhaps these artists could identify with the nationality of their birthplace as well as with their American or English nationality. Accordingly, these maps complicate the notion of "American" or "English." Even with regard to the CMOA collection itself, while it may feel overwhelmingly American, these maps indicate that the "American" classification could encompass other nationalities as well.





Click here for details on how these visualizations were created.

ABOUT

About the Project

The Carnegie Museum of Art (CMOA) is a contemporary art museum in Pittsburgh, Pennsylvania. Founded in 1895, the museum calls itself the first museum of contemporary art in the United States, seeking the "Old Masters of tomorrow." The museum features over 30,000 objects covering a broad range of medium and form, including painting, sculpture, photographs, film, and digital imagery.

Key to CMOA's aspirations of collecting the Old Masters of tomorrow is understanding contemporary art as a contemporary issue that engages with current social events and conditions. In a note from their director, CMOA asserts that their programming, exhibits, and publications "frequently explore the role of art and artists in confronting key social issues of our time, combining and juxtaposing local and global perspectives." Their mission statement champions creativity with a global focus, reading in full:

> *We create experiences that connect people to art, ideas, and one another. At CMOA, we believe creativity is a defining human characteristic to which everyone should have access. CMOA collects, preserves, and presents artworks from around the world to inspire, sustain, and provoke discussion, and to engage and reflect multiple audiences.*

This project takes its lead from CMOA's mission statement, examining their accession records to get a sense of how global their collection is. CMOA recorded the nationality of artists within each artwork's accession record, and was fairly thorough in this effort. In the 4,869 accession records for their contemporary art department, only 11 have artists whose nationality is classified as "undetermined." While there is no silver bullet for determining the diversity of a museum's collection,

the data visualizations in this project will tease apart aspects of CMOA's contemporary art collection to get a better understanding of the nationalities and countries represented in their artworks.

About the Dataset

As part of their 120th anniversary, CMOA released collection records for all of their accessioned works. Each record includes information about the artwork (including title, creation date, and date acquired) as well as information about the artist (including name, nationality, and birthplace).
Since these are working records, the data contributors have a vested interest in maintaining thorough and accurate records. However, human error is very much a possibility within any dataset, and CMOA notes that the dataset may contain incomplete data or errors. Further, the dataset is not a static repository, but a continual project. Art historical research is ongoing at the museum, so CMOA advises downloading the most current version of the dataset to benefit from any updates to the records.

Technical Decisions

I primarily used Tableau to create data visualizations for this project. I wanted to develop expertise in Tableau, so I used the software to create different types of graphs. Tableau, however, was not a good fit for creating timelines, so I used TimelineJS for the timeline in the *Items on View* section.
.
With regard to the design of the site itself, I wanted it to be as clear and legible as possible. Mobirise offered an intuitive and simple design. I structured the homepage so that a link to the "About" page is prominently displayed in a bright red button, to encourage users to visit this page first to have more context for the visualizations and the project overall. Otherwise, I gave the four avenues into exploring nationality within the collection more or less equal footing, allowing users to click on what interests them most as opposed to forcing a certain order with a scrolling screen. I believe giving users the freedom to choose what interests them most makes them feel more engaged with the project. Further, the four aspects I explore are not necessarily linear, so forcing an order does little to enhance the understanding of the data visualizations themselves.

Data Decisions

While data and data visualizations can be useful for spotting trends, data alone cannot represent an objective truth. Everything about data is subjective—data collection involves a subjective choice as to what you believe is valuable, and can involve errors and oversights. Data visualization software often requires data to fit a certain mold, eliminating nuance and outliers during the data cleaning process. In order to provide more context and transparency, I have included below specific steps and decisions I made when cleaning the data so that I could make legible visualizations.

With regard to the CMOA dataset overall, records were classified into six departments: contemporary art, decorative arts and design, film and video, fine arts, Heinz Architectural Center, and photography. I chose to focus on art that fell under the contemporary art department. From a practical standpoint, winnowing the dataset down from 28,154 records was a necessity, as I knew my data cleaning would involve some detailed record-level work. The contemporary art department has 4,869 records, which was a more reasonable amount for the 10-week period I had to work on the project. As to which department to choose, I selected contemporary art as it felt the more general, and thus inclusive, than some of the other more specialized departments (like film and video or photography). Further, the department felt more in line with CMOA's point of pride—being the first museum of contemporary art museum in the United States—than some of the other departments, like fine arts.

MAPPING ARTISTS

The dataset provides the nationality of the artists for the artworks. To prepare the data for mapping, I had to change these descriptions to their country name (for example, change "Finnish" to "Finland"). Within OpenRefine, I identified spelling patterns across the nationalities (such as those ending with "ish" or "an"), used chomp GREL statements to trim these endings, and then replaced them with the appropriate endings.

Some artworks had multiple artists, and thus multiple nationalities. I separated the nationality column by separator, and focused on cleaning the nationality of the first artist only. While not ideal, including the nationalities of multiple artists for one piece would have unequally weighted the nationalities, framing the data in a way that equates one artwork by four artists to be the same as four separate artworks. Fortunately, in general multiple people tied to a piece were of the same nationality.

Additionally, the dataset lists both British and English, Scottish, and Welsh as nationalities. I would have liked to have parsed out the British entries to include more specificity; however, Tableau's mapping function only has United Kingdom as a country, and does not map specifically for England, Scotland, and Wales.

ACQUISITION TRENDS

The "credit line" field in the dataset specifies how the artwork was acquired. The credit lines were very detailed, with names of specific donors. Within OpenRefine, I was able to use text filters on keywords like "gift of" and then run GREL replace statements to clean the data to just four categories: purchase; gift; partial gift, partial purchase; and museum appropriation.

The dataset also provided creation and acquisition dates for the artworks as a text string data type. Within OpenRefine I transformed these entries into the date data type, and then ran a GREL formula to subtract the creation date from the acquisition date and return a year amount.

Because American artworks comprise over 54% of the dataset, I decided to break these visualizations down by American and non-American art. While it would have been interesting to plot every nationality to see more detail and nuance, too many nationalities have too few artworks to make those graphs readable. Of the 56 nationalities recorded, 21 of them have 3 or less artworks. For the visualization of the time between an artwork's creation and acquisition, then, plotting each country would have yielded dozens of small squiggles that do not clearly show purchasing trends. The graphs with American art versus art from outside the United States are able to tell a clearer story.

Some date entries had circa, or a range of years. In order to map dates, I had to normalize them, so when there was a range I took the average of the dates. I put a 20-year max on taking the average of date ranges. Ranges that exceeded 20 years were too broad to be meaningful; for example, a handful had creation dates like 1916-1956 or 20th century. I excluded artworks with these overly broad and general date ranges, as I would not be able to map them and there were relatively few (approximately 25 of the 4,689 entries).

ITEMS ON VIEW

The dataset provides the physical location of each artwork, specifying whether it is in a specific gallery, the museum lobby, the museum grounds, or not on view. I standardized these in OpenRefine, to be either "on view" or "not on view" to create the visualizations.

The dataset did not specify how long the piece has been in a location, so it is possible that the pieces move frequently or stay in their location for years. It would have been nice to have this information to situate the data more. Further, one of the categories was “off-site.” The dataset does not have a data dictionary, so it is not clear if the piece was off-site for restoration or off-site as part of a loan to another museum. Either way, I chose to classify artworks with this classification as “not on view,” because CMOA was not themselves putting that piece on view.

For the timeline, it was tricky to identify the nationalities of the 32 artists and artist collectives that participated in the Carnegie International, 57th Edition. The other artists in the timeline were either in the CMOA dataset, or their place of residence and birth were the same. This was not always the case with the artists from the Carnegie International. Accordingly, I referred to the documentation that CMOA provided about the event, which listed all of the "national affiliations by residence and birth." Of the 32 artists from the event, 20 lived in the United States and 26 nations were represented in country of birth. Recording all of these nationalities differs from how artist nationality was recorded within the dataset, as only one nationality was chosen. However, as the dataset did not clarify how nationality was chosen (specifically, if residence was favored over place of birth, and how residences in different countries were weighted), the best option was to include the nationalities as described by CMOA in their event programming.

THE AMBIGUITY OF NATIONALITY

The dataset provides the birthplace of the artists as one long text string (for example, “Seattle (King County, Washington, United States)”). Within OpenRefine, I clustered these birthplaces to standardize any variants. Then, I used GREL replace statements to transform the birthplaces into a string that used a comma as a delimiter (for example, “Seattle,Washington,United States”). Since Tableau maps through a hierarchy—city, state, country—I parsed these values out by separating the columns by the comma delimiter.

Often times the dataset only recorded the home country of artists. Despite research, there were six artists for whom I could not track down their specific birthplace. Either this information was not available online or their town was not appearing on a map. Since there were so few artists out of the 1,304 total that had this issue, I omitted them from the visualization.

References

"About." n.d. Carnegie Museum of Art. Accessed December 2, 2019. https://cmoa.org/about/.

"Art & Architecture Thesaurus Full Record Display (Getty Research)." n.d. Accessed December 3, 2019. http://www.getty.edu/vow/AATFullDisplay?find=appropriation&logic=AND&note=&english=N&prev_page=1&subjectid=300311629.

Burleigh, Paula. "The Politics of Museum Joy." *Art Journal* 78, no. 2 (March 2019): 125–28. https://doi.org/10.1080/00043249.2019.1626164.

Carnegie Museum of Art. (2015) 2019. "Cmoa/Collection," September. https://doi.org/10.5281/zenodo.35013.

Drucker, Johanna. "Graphical Approaches to the Digital Humanities." *Digital Humanities Quarterly* 5, no. 1 (2011).

"Exhibitions." n.d. Carnegie Museum of Art. Accessed December 5, 2019. https://cmoa.org/exhibitions/.

Groskopf, Christopher. "The Quartz Guide to Bad Data." Quartz. December 15, 2015. https://qz.com/572338/the-quartz-guide-to-bad-data/.

"History." n.d. Carnegie Museum of Art. Accessed December 2, 2019. https://cmoa.org/about/history/.

McNamee, Donald. 1983. "Abstract Painting and Sculpture in America 1927-1944." *The Structurist* 0 (23): 102.

"Participants." Carnegie International 2018. Accessed December 9, 2019. https://2018.carnegieinternational.org/participants/.

Ramos, E. Carmen. "The Latino Presence in American Art." *American Art* 26, no. 2 (2012): 7–13.

Rawson, Katie, and Trevor Muñoz. "Against Cleaning." Curating Menus, July 6, 2016. http://curatingmenus.org/articles/against-cleaning/.

"Past Exhibitions." n.d. Carnegie Museum of Art. Accessed December 5, 2019. https://cmoa.org/exhibitions/past/.

Elective Coursework

Data Curation and Policy // IS 262B
Jillian Wallis
Spring 2019

Abstract
A critical component of research is data management. Responsible data management benefits research on numerous fronts—it promotes trust in the data; it facilitates data reuse; and it provides standards and context that support the longevity of a project, which can be important if the research project is a years-long endeavor with a multitude of changing contributors. Data management, however, can be complicated in interdisciplinary research projects, as different disciplines have different data standards and practices. This is where the expertise of those in the information studies field comes in. Professionals with a background in information studies are specifically trained to think about preservation, trust, and access of information.

We collaborated with Dr. Kara Cooney, a professor in the Department of Near Eastern Languages and Cultures. Her research explores coffin reuse in Ancient Egypt. This work is an expansive endeavor, involving eight years of data collection. All in all, Dr. Cooney has examined over 300 coffins in over 20 museums and private collections around the world, generating over 100,000 photographs as well as qualitative data on each coffin, totalling one terabyte of data.

Dr. Cooney's work is interdisciplinary, drawing on art history, archaeology, and Egyptology. Her data also has rights' concerns, as museums have different permission controls with regard to disseminating the photographs of coffins. With regard to transparency, appropriate measures are needed as her data is qualitative, consisting of Dr. Cooney's observations of signs of coffin reuse. And finally, Dr. Cooney works with graduate students who were not necessarily there since the beginning of her research, which can affect the longevity of her data.

We created a data management plan that supports these concerns, and offers a path forward for sharing her findings online in order to facilitate data reuse. Our recommendations cover a number of areas that promote data transparency, reuse, and longevity, including description, rights, storage, repositories, citation, budgets, and timelines.

Inclusion in portfolio
I included this report in my portfolio because it shows a deep understanding of data lifecycles, metadata, transparency, trust, and preservation—all of which are critical concepts with applications in a variety of information domains. Further, this project spanned two terms; in 262A we created the data management plan, and in 262B course we implemented the plan. Accordingly, I was able to go more in depth with the report and with the client to deliver the most useful recommendations possible.

Data Management Plan: Implementation Report
Dr. Kara Cooney | Coffin Reuse Research

Prepared by
Savannah Lake, Ashton Prigge, and Marisa Purcell
IS 262B | Dr. Jillian Wallis
June 12, 2019

Table of Contents

I. Overview

Dr. Kara Cooney is a professor of Egyptian art and architecture at UCLA, as well as the chair of the Department of Near Eastern Languages & Cultures. Specializing in craft production, coffin studies, and economies in Ancient Egypt, Dr. Cooney has extensively researched Ancient Egyptian funerary practices, contextual architecture, funerary arts, and material culture.

Her current research explores coffin reuse in Ancient Egypt. For Ancient Egyptians, coffins were an integral part of the afterlife, facilitating the transformation of the dead. However, during the 21st Dynasty (circa 1150 BCE), Egypt and its neighboring civilizations in the Mediterranean and the Near East experienced socioeconomic and political instability, afflicted with drought, famine, and foreign invasion.[1] As a result, trade networks collapsed, and Egypt no longer had access to the materials necessary to create coffins. As coffins were not believed to protect the dead in the long term (instead seen as enabling rebirth immediately after death), Ancient Egyptians resorted to reusing older coffins in order to ensure safe passage to the afterlife for the recently deceased.[2] Dr. Cooney's research revolves around evaluating these coffins for signs of reuse, looking for modifications in decoration, names, and coffin parts.

Dr. Cooney's research on coffin reuse is an expansive endeavor, involving eight years of data collection. All in all, Dr. Cooney has examined over 300 coffins in over 20 museums and private collections around the world, generating over 100,000 photographs as well as qualitative data on each coffin. This qualitative data was collected and saved as PDF field note files, which were then manually entered into an Excel database. The photos and field notes total one terabyte

[1] "Update from ARCE: Current Research, Excavation and Conservation Projects in Egypt," *NILE* Magazine, October-November 2018, 59.
[2] "Update from ARCE," *NILE* Magazine, 59.

of data. Dr. Cooney is no longer collecting data. At this stage of her research she is interested in cleaning, curating, and sharing her data via publications and her research and educational website called WikiArtifact.

At the start of this quarter, we met with Dr. Cooney and her research team to review our data management plan from the winter quarter to determine next steps for data management and WikiArtifact. Dr. Cooney expressed concern about the financial sustainability of WikiArtifact, and wanted to ensure WikiArtifact can be financed and maintained for the long term. Last quarter, Dr. Cooney envisioned WikiArtifact as a visual, interactive online database of her research data, complete with visual tags that users could click on to view and search multiple data elements of the coffins, such as location, mythical imagery, and coffin materials. Dr. Cooney wanted WikiArtifact to be accessible, usable, and dynamic, with 3D imaging of her coffins and easy sharing via social media. The database would be collaborative, with vetted researchers adding their coffin-related data.

Accordingly, we recommended Omeka as a platform that could accommodate all of these functionalities. The platform is available via two routes: Omeka.org and Omeka.net. Omeka.org is a free, open-source platform, but requires the user to have their own server space.[3] Maintaining a server can be complicated and costly, so we investigated Omeka.net. With Omeka.net, users pay for server space hosted by Omeka based on their storage needs.[4] Omeka's support team gave us a quote of $5000 a year based on Dr. Cooney's need to store one terabyte of data. Even if Dr. Cooney were able to eliminate duplicate photos and reduce her storage needs to half as much (500 gigabytes), Omeka would cost $3000 a year.[5] Though Omeka checked several boxes off of

[3] "Omeka Classic," https://omeka.org/classic/. (Accessed June 6, 2019).
[4] "Pricing," Omeka.Net, https://www.omeka.net/signup. (Accessed June 6, 2019).
[5] "Omeka.Net Price List," 2018, Corporation for Digital Scholarship.

Dr. Cooney's must-have list, it proved to be too costly and the desired qualities of WikiArtifact needed to be reevaluated.

Accordingly, instead of envisioning WikiArtifact as a visual database, Dr. Cooney expressed interest in sharing her data through photo essays.[6] Dr. Cooney still envisions the project as being collaborative, and would like it to be shareable through social media. We took these concerns into consideration as part of this project, and looked into how to facilitate more cost-effective WordPress and Drupal sites for WikiArtifact (see the section entitled "Objective 2: Identify Server and Platform Options for WikiArtifact," starting on page 15).

II. Project Objectives

Through our conversations with Dr. Cooney and her research team, we identified two objectives for this project that will best prepare Dr. Cooney's data for the creation of WikiArtifact. First, we will clean the data in Dr. Cooney's Excel database. Currently, data is difficult to extract from the spreadsheet, in large part due to an inconsistent and unwieldy metadata schema as well as a lack of controlled vocabulary. A more strategic, consistent approach to metadata schemata and controlled vocabularies will correct these issues, and get the data in a more contextualized, standardized state. This will both aid in current use of the spreadsheet and it will prepare the data for safe upload into WikiArtifact.

Our second objective is to identify server options and platforms for WikiArtifact. Dr. Cooney and her team were very interested in understanding the technical options for WikiArtifact. Our research and recommendations for server and platform options will prioritize

[6] Kara Cooney, Interview with Dr. Kara Cooney and her research team, In-person, April 18, 2019.

the long-term stability of the data, funding limitations, as well as Dr. Cooney's vision for the project.

III. Objective 1: Cleaning the Excel database

The current Excel database houses all of the qualitative data Dr. Cooney has collected on 300 coffins throughout the world. While robust and largely functional, Dr. Cooney's research team has encountered problems when trying to extract data from the spreadsheet. We have identified two elements of the spreadsheet—the metadata schema and the controlled vocabulary—that could be expanded upon and standardized to streamline the spreadsheet, making it more navigable and usable.

A. Metadata Schema

The Excel database currently uses a homegrown metadata schema, tailored to the specific needs of Dr. Cooney's research. The schema is comprised of 15 fields, including information on the holding institution of the coffin (city, museum, accession number), descriptive metadata about the coffin (coffin type, coffin part, dating, provenance, Niwinski number, name(s) of the deceased, title), reuse information about the coffin (date examined, reuse score, type(s) of reuse), and miscellaneous notes (notes and other notes). These fields set a strong foundation for us to work off of. Ultimately, we created a revised schema that more accurately captures the complexity of Dr. Cooney's data, as well as created a crosswalk to Dublin Core to enable potential data sharing on a larger scale.

a. Revised metadata schema

We identified three issues in Dr. Cooney's homegrown schema that we sought to address with a revised schema: inconsistency, a lack of specificity, and insufficient description. When reviewing the spreadsheet, we saw that metadata fields were used inconsistently in some cases. The "provenance" field, for example, sometimes included the museum the coffin was housed at, the name of the excavation team, and information about the buyer and seller of the coffin.[7] This inconsistent entry is in part because the field of "provenance" is too broad. Many elements contribute to an item's provenance, include acquisition details, creation information, and the item's holding information. Broad terms ellide this specificity. This complicates data retrieval, as it is not clear what information will be found in each field. Further, during the initial stages of data collection, having more specific, granular metadata fields encourages more thorough and consistent data collection. In addition to the overly broad metadata fields that are inconsistently used, we found that Dr. Cooney's spreadsheet lacked some critical metadata fields, such as administrative data on access rights or provenance metadata for the data itself (not the coffin).

Ultimately, we wanted to create a metadata schema that was useful to Dr. Cooney, and streamlined features of her existing database. Our fully revised scheme can be found in the appendix on page 30. Our implementation of the schema on a set of data can be found in the appendix on page 37. As noted on both of these pages, the new scheme provides for provenance, descriptive, and administrative metadata. We have also defined each element in the revised metadata schema, so Dr. Cooney and her team can fully understand the new schema (appendix, page 30). Starred elements on this appendix item represent Dr. Cooney's initial elements that we

[7] Kara Cooney and Amber Wells, 2018, "Coffin List (FULL) 3.0."

carried forward to the revised schema. The other elements that are not starred, represent revisions and additions we made.

First, our revised scheme breaks apart several of Dr. Cooney's original fields in order to capture the nuance and complexity of the data. This granularity improves retrievability and entry consistency, and follows the principle of "atomizing" information—essentially, breaking up information so that each metadata field only contains one type of data.[8] An example of this can be found with the revisions we made to the "dating" field. Previously, Dr. Cooney's schema only had one field for describing the date of the coffins. This meant that the field was often filled with long, continuous blocks of texts, such as "19th dynasty to mid 20th dynasty ???"[9] In the revised schema, the "dating" field has been broken into five distinct fields: coffin start time period and coffin end time period, both accompanied by fields to add period descriptors (such as early, mid, late), as well as field to denote ambiguity about the period. From the above example of "19th dynasty to mid 20th dynasty ???", the information would be "atomized" into distinct chunks: 19th dynasty, mid, 20th dynasty, and ambiguous. Breaking "dating" into these five categories makes the Excel noticeably more navigable, as you can sort and filter for these facets by column.

Similarly, we broke out the "provenance" field to be more specific, adding fields such as "buyer," "seller," and "acquisition date." We also added a few categories for data provenance, including "data collector" and "reuse observations and explanation." These fields were missing from the previous spreadsheet, but are important for tracking the provenance of the data, thereby engendering trust in the data for external researchers who access WikiArtifact and wish to reuse the data.

---

[8] Carly Strasser, 2015, "Research Data Management," NISO Primer Series, Baltimore, MD: National Information Standards Organization, 5.
[9] Cooney and Wells, "Coffin List (FULL) 3.0."

The final two fields from Dr. Cooney's initial schema that we reworked were "notes" and "other notes." These sections functioned as catch-all fields within the database for miscellaneous information related to the coffins. We reviewed the content of these fields, and identified three commonalities: notes on file location, information on related coffins, and reuse observations for the coffin. We thus eliminated these two categories, which were vague and inconsistently utilized, and created these three new metadata categories. It is advised, however, that Dr. Cooney and her research team do their own review of the "notes" and "other notes" sections, to determine if there was any other categories information that should be added to the revised metadata schema. We are neither Egyptologists nor were we deeply entrenched in the data collection process. Dr. Cooney and her team may be able to identify additional patterns and commonalities within these two sections.

b. Dublin Core Crosswalk

In addition to revising Dr. Cooney's metadata schema to make the data easily retrievable and navigable, we also wanted to provide Dr. Cooney with the option to make her data more interoperable via a standardized schema. Last quarter, we suggested reviewing two potential standardized schemata for Dr. Cooney's data: MIDAS Heritage and Dublin Core. Developed by the Forum on Information Standards in Heritage and recommended by the Digital Curation Centre, MIDAS Heritage is a robust metadata standard for describing archaeological buildings, sites, and artifacts.[10] When reviewing the MIDAS Heritage standard, however, we found their categories and subcategories too detailed, and perhaps overwhelming to a research team with limited resources for information management.

[10] "MIDAS Heritage: The UK Historic Environment Data Standard," 2012, Forum on Information Standards in Heritage, https://historicengland.org.uk/images-books/publications/midas-heritage/midas-heritage-2012-v1_1/, 22 and 27.

Conversely, the Dublin Core metadata standard is a simple, low-cost metadata standard for digital objects. The schema was "designed to be extremely simple, flexible, and extensible" to encourage as wide adoption as possible. Dublin Core is comprised of just fifteen core elements, which are all optional and repeatable.[11] After evaluating its core elements, we determined that Dr. Cooney and her team would feel comfortable with the schema, especially as compared to other standardized schemata. Further, because the basic elements are simple and flexible, a wide variety of communities are more likely to use it—making the schema a good fit for Dr. Cooney's data, which spans the disciplines of art history, Egyptology, and archaeology.

Our Dublin Core crosswalk can be found in the appendix, on page 33. Should Dr. Cooney ever wish to make her data more shareable or interoperable, she now has a clear roadmap for doing so. Further, Dr. Cooney could encode these Dublin Core elements into the back-end of the WikiArtifact site—not the front-end—so that this metadata is searchable and retrievable without ruining the aesthetic or preferred term usage for metadata fields within the photo essays on WikiArtifact (for example, if Dr. Cooney would like to maintain the metadata field "types of reuse" instead of the more generic Dublin Core field "subject").

We encountered a few roadblocks when mapping Dr. Cooney's revised schema to Dublin Core. Overall, there were instances of unclear mapping. More specifically, we often found that we were unclear about whether we were describing the coffin, the photo of the coffin, or the dataset about the coffin. This problem arose in mapping to Dublin Core fields such as "contributor", "date," "type", and "language." Generally speaking, we prioritized describing the coffin, not the image of the coffin or the dataset. However, for "contributor," for example, we

[11] Stephen J. Miller, 2011, *Metadata for Digital Collections: A How-to-Do-It Manual*, London, UK: Facet Publishing, 51.

felt it was important to list Dr. Cooney as a "contributor" for her role in data collection, even though we were technically describing the coffin, not the data set about the coffin. While somewhat inconsistent, omitting this metadata would erase a lot of the context for the coffin within the setting of WikiArtifact. Similarly, the "type" of object per Dublin Core could be a physical object (the coffin itself) or a data set. The "language," too, could refer to the language on the coffins, or the language of the dataset. This ambiguity is a natural consequence of using Dublin Core, whose simplicity does not allow for nuanced description that could delineate these relationships. Nevertheless, the strengths of Dublin Core with regard to wide-scale adoption, simplicity, and interoperability make it the best fit for Dr. Cooney's data.

Additionally, some fields from Dr. Cooney's schema mapped onto several Dublin Core elements. For example, "name of the deceased" maps onto both "subject" and "description." Per Dublin Core's usage guideline, "subject" refers to "the topic of the content of the resource," often described in keywords or key phrases, while the "description" field is "an account of the content of the resource," serving as "a potentially rich source of indexable terms" that can use full sentences.[12] The "name of the deceased" sits between these two categories, without a clear-cut home.

It was also difficult to create the crosswalk without knowing the new context in which the data would be used. When crosswalking, it is not always necessary to map every element from the old schema to the new schema. It is only necessary to map the elements that are relevant to the new context. For example, the category "reuse score" would not always need to be mapped onto "description," if the new context in which the data is being used is not concerned with this

[12] "DCMI: Using Dublin Core," http://www.dublincore.org/specifications/dublin-core/usageguide/elements/. (Accessed April 27, 2019).

particular metric. In mapping the schema, we tried our best to be as inclusive as possible, to account for whatever new contexts the data may be used in.

There were also a few Dublin Core elements that we were not able to map to, as Dr. Cooney had not collected data on that front. "Format," for example, was unmappable, as Dr. Cooney did not collect data on dimensions of coffins or materials. Since Dublin Core does not require usage of each element, this is not a serious hindrance. But it does show how schemata are not always easily matched or aligned.

Finally, in completing the crosswalk we noticed that some categories from Dr. Cooney's schema do not exist in Dublin Core. This meant that some data collected is lost in the crosswalk to Dublin Core. For example, the location of the holding institution for the coffin, which in Dublin Core would be the location of the "publisher," did not make it during the mapping process. Fortunately, none of the affected elements were especially significant to understanding Dr. Cooney's research.

All of these issues are endemic to mapping and Dublin Core in general.[13] Despite this, the benefits of potential widespread sharing and interoperability make mapping a valuable exercise and potential option for Dr. Cooney's data.

B. Controlled Vocabulary

Dr. Cooney's Excel database suffers from inconsistent naming practices and data entry, in part due to turnover in research assistants and in part due to the nature of her data, which is subjective and thus can generate ambiguous descriptors. This has made it difficult to

[13] Mary S. Woodley, 2016, "Setting the Stage," In *Introduction to Metadata*, edited by Murtha Baca, http://www.getty.edu/publications/intrometadata/metadata-matters/.

systematically and efficiently extract information from the database, since each column had so many variant terms that filtering columns did not always retrieve accurate or complete results. Often when research assistants were attempting to create charts from the data, they had to manually review cells.

We took several steps to standardize vocabulary use within the spreadsheet. First, we collocated the terms. We ran the entire Excel spreadsheet through OpenRefine, a data cleaning application. This allowed us to collocate terms, and determine authority terms amongst variant terms. This provided a holistic view of what sort of terms were used and where the variation took place. We completed this for categories that required little Egyptology expertise, and have shared a list with Dr. Cooney and her research team to review and approve. However, there were quite a few categories that would require Egyptology expertise to understand the variant terms and their relationships. Accordingly, Dr. Cooney's research team will need to review these categories, and determine authority terms directly. Dr. Cooney has communicated that this will likely happen in Fall 2019, when they have more graduate student researcher support. This is a critical component of the data clean-up, and should be prioritized.

Our next step toward standardizing Dr. Cooney's vocabulary was to integrate the Getty Vocabularies, where applicable. Although not a perfect fit for her very specific data regarding coffin reuse, there are a number of fields within her spreadsheet that have been standardized via Getty Vocabularies. The Thesaurus of Geographic Names could be used for locations, while the Union List of Artist Names could be used for museum names. A full list of these recommendations, as well as other style and naming conventions, can be found in the appendix on page 35. Integrating the Getty Vocabularies into the spreadsheet will make Dr. Cooney's data

more interoperable; however, this is not as high of a priority as cleaning up the variant terms via OpenRefine or migrating the spreadsheet to the revised metadata schema, as the Getty Vocabulary standardization will only help the external researchers who are using these vocabularies. The other two data standardization tasks will help all users, including Dr. Cooney and her team. Accordingly, Dr. Cooney should only implement this recommendation if she has the time and resources to do so.

Finally, we found that our revised metadata schema solved some of our vocabulary issues, particularly with regard to ambiguity. The coffin dates, for example, are now broken into five columns that are granular enough to avoid variance. Where before, it was common to have variant dates like "early to mid 21st Dynasty," "early-mid 21st Dynasty," and "early/middle 21st Dynasty," now, the five metadata fields in the revised metadata schema encourage more consistent and standard inputs.[14]

C. Workflow and Implementation Recommendations

In order to jumpstart the database clean-up, we have revised a subset of Dr. Cooney's data with both the new metadata schema as well as our controlled vocabulary recommendations. These entries will also serve as an example of clean data, which Dr. Cooney and her team can review and refer to during their data cleaning process.

Dr. Cooney's team selected the coffins from Museo Egizio in Turin, Italy as the ideal starting point for WikiArtifact as they believe that Museo Egizio will be the most flexible in

[14] Cooney and Wells, "Coffin List (FULL) 3.0."

terms of image and data sharing.[15] Accordingly, we transformed the coffin entries from this museum, devising the following workflow for data-cleanup:

1) Migrate data from the old metadata schema to the revised schema.

a. Review and understand the new metadata schema. Read the definitions of the new metadata fields (page 30).
b. Understand the differences between the old schema and the revised schema.
c. Create a new Excel spreadsheet, with the new metadata fields as the header.
d. Copy and paste data from the old metadata spreadsheet to the new spreadsheet with the revised schema, moving data to the new metadata fields.

2) Standardize data through controlled vocabularies.

a. Import the spreadsheet into OpenRefine.
b. Correct for variant terms: select a column and then filter by text facet. On the left side panel, all variations within the column will appear. Cluster the terms, and then enter your preferred authority term. Merge and recluster the items.
c. Integrate the Getty vocabularies: review which columns should use the Getty vocabularies (page 35). Add these terms to the spreadsheet. For Getty terms that can be applied to multiple cells, you can apply them at scale in OpenRefine through clustering and assigning a preferred authority term again.

3) Clean the data. In addition to clustering variant terms, OpenRefine is a powerful tool for editing and transforming data.

[15] Cooney, Interview with Dr. Kara Cooney and her research team.

a. Eliminate white space. White space is extra spacing within a cell that is invisible to the eye, but can cause problems in data curation. Best practice for cleaning data is to eliminate white space.
b. Clean up the "types of reuse" field. In OpenRefine, you can break apart cells into multiple rows. This will allow for better manipulation, sorting, and filtering of this column, improving data retrieval. These edits do not export well into Excel, so this search functionality may be best done exclusively within OpenRefine.

We encountered a few difficulties when carrying out the data transformation for these 30 entries, which should be noted so that Dr. Cooney and her team can do their best to avoid them. During the initial steps of moving data from the old metadata schema to their new fields, we found the Excel spreadsheet to be somewhat cumbersome and not user friendly due to the sheer number of columns. We would recommend freezing the header so that team members do not need to scroll up to remember each column name. Column order could also be adjusted to best suit the team members' workflows. With regard to the controlled vocabularies, the Getty vocabularies rely on hierarchies to mark relationships. This may not be the most intuitive structure for new users. Fortunately, the two Getty vocabularies we are recommended—Union List of Artist Names (ULAN) and Thesaurus of Geographic Names (TGN)—are more straightforward on this front than a vocabulary like the Getty Art & Architecture Thesaurus, whose subject and topics are highly interconnected and more hierarchical than names and locations.

As for implementing the above workflow, we would highly recommend that Dr. Cooney and her team attend a training on OpenRefine. Utilizing OpenRefine as part of their data

clean-up will save Dr. Cooney and her research team an immense amount of time as the program is intuitive, powerful, and can transform data on a large scale. The Data Science Center at UCLA Library regularly hosts workshops on OpenRefine.[16] We would recommend attending one of these workshops or contacting the Data Science Center directly for training.

As described in the previous data management plan, data stewardship is an active, ongoing responsibility. The initial clean-up and transformation of data does not signal the end of data management practices. There are a few fields that are likely to change throughout the lifetime of the data, including "file location notes" and "access rights." Any changes to the data on these fronts should also happen to the Excel database. In fact, it is recommended that researchers revisit all data management documentation, including the previous plan and this report, on a weekly basis, to ensure follow-through and consistency as well as to record any updates.[17]

IV. Objective 2: Identify server options and platforms for WikiArtifact

A. UCLA IT Options

As addressed above, new server and platform options were needed in order to cut down on cost and fit the WikiArtifact's new format as photo essays. UCLA's IT department offered three options to build and host a website.

1. UCLA IT Option #1:

The most cost-effective option would be for Dr. Cooney to create her own WordPress site that could be coupled or uncoupled with a CPanel hosted by UCLA's IT department. A CPanel is

[16] "UCLA Library Events," UCLA Library, https://www.library.ucla.edu/events/data-cleaning-openrefine. (Accessed June 8, 2019).
[17] Strasser, "Research Data Management," 5.

a "web based hosting control panel provided by many hosting providers to website owners allowing them to manage their websites from a web based interface. This program gives users a graphical interface from which they can control their portion of the... server."[18] Hosting their own website on the CPanel means that Dr. Cooney's team would be responsible for maintaining the server; as such, they would be responsible for any security threats that UCLA deems concerning. This also means that Dr. Cooney and her team would need to use a platform like Drupal or WordPress for the site, and potentially pay a developer to add and maintain any plugins they want the site to have. Though this server option is the most cost effective at $28 a month, the responsibility of maintaining their CPanel could take up valuable resources like the time of graduate student researchers and funding for a developer to perform security maintenance.[19]

2. UCLA IT Option #2:

The second option UCLA IT offers is a service called Site Factory.[20] This option is $300 a month but comes with much more than server space.[21] Site Factory offers templates to create an interactive website, and UCLA's IT department used Site Factory to create several websites.[22] The style guide and templates use Drupal.[23] This option provides for a fast launch: through Site Factory, websites can be launched within 8 hours, versus 180 hours when creating your own custom site on cPanel. The downside of the Site Factory option is that it is not as customizable as

[18] "What Is CPanel? How to Use CPanel for WordPress Hosting," WPBeginner, https://www.wpbeginner.com/glossary/cpanel/. (Accessed June 11, 2019).
[19] Damon Wolf, Interview with Technical Account Manager | Information Technology Services at the University of California, Los Angeles, Phone, April 17, 2019.
[20] Ibid.
[21] Ibid.
[22] Ibid.
[23] Ibid.

Dr. Cooney may hope for. Additionally, this option is more costly per month and would require hiring a Drupal developer for roughly $110 an hour to develop the site and potentially maintain the site depending on the level of comfort Dr. Cooney and her team.[24]

3. UCLA IT Option #3:

The third option would be for Dr. Cooney and her team to manage their own server. To run a platform like Omeka with several plugins on a dedicated server, the cost would be nearly $200 per month for the server, and $0.09 for every GB of image storage (1TB would therefore be around $90/month).[25] This does not include the need for a web developer to create the site. This option is not recommended because it can be time-intensive, costly, and require some expertise.

B. Server and Platform Options Based on Dr. Cooney's New WikiArtifact Vision

In speaking with UCLA's IT department and Dr. Cooney, as well as through researching platforms, it became clear that the vision of WikiArtificat had to change based on the resources available to Dr. Cooney. All server and platform options had trade-offs such as functionality, budget, or sustainability. Dr. Cooney and her team made it clear that sustainability and usability are the top priorities.[26] After understanding the budget needed to support running a large amount of data on a server as well as customized APIs, Dr. Cooney and her team decided WikiArtifact should be a more streamlined site formatted as photo essays.[27] They also determined that WordPress is a good option to host WikiArtificat because Dr. Cooney's team has used and is comfortable with the platform.[28] Our team balanced the priorities of budget, accessibility, and

[24] Ibid.
[25] Damon Wolf, Interview with Technical Account Manager | Information Technology Services at the University of California. .
[26] Cooney, Interview with Dr. Kara Cooney and her research team.
[27] Ibid.
[28] Ibid.

sustainability, ultimately deciding that WikiArtifact should be built using WordPress, and Dr. Cooney should use WordPress server space to host the site.

WordPress offers a business option for $25 a month with unlimited storage space, which is necessary for Dr. Cooney's terabyte of photographs.[29] This pricing structure means a WordPress platform and hosting option would cost $300 year, a far cry from Omeka's $5000 a year. The pricing of WordPress combined with the team's general familiarity makes WordPress the most sustainable option for WikiArtifact. Furthermore, WordPress features that come with the business plan will be helpful to Dr. Cooney and her team, including social media compatibility, Google Analytics, live chat setup support, unlimited premium themes, and the option to install customized themes and plugins.[30] If Dr. Cooney does decide she wants to customize her WordPress site beyond the templates that WordPress offers, she may need to consider putting funding aside for a developer. Lastly, Dr. Cooney could start with a smaller and less costly WordPress site, and upgrade the site in the future when she uploads the full terabyte of data.[31]

V. Roadblocks and Future Recommendations

We have identified potential roadblocks on the data management side as well as with the technical implementation of WikiArtifact. Proactive and thoughtful stewardship will go a long way in addressing these issues.

A. Data Management

[29] "WordPress.Com Plans and Pricing – Get Started for Free Today!," *WordPress.Com* (blog), February 23, 2016, https://wordpress.com/pricing/.
[30] Ibid.
[31] Ibid.

a. Photo permissions

During our meeting with Dr. Cooney and her research team at the start of the quarter, Dr. Cooney communicated that museums can be protective around the dissemination of images and information about their holdings.[32] While Dr. Cooney took photos of the coffins personally and thus technically has copyright over that creative work, Dr. Cooney is very sensitive to museum needs and wants to maintain good working relationships with them. Accordingly, Dr. Cooney and her team will need to secure permissions from each museum in order to distribute coffin photos on a public website like WikiArtifact. Dr. Cooney could also watermark the photographs to prevent unauthorized dissemination, but securing photo permissions is the highest priority.

Dr. Cooney indicated that some museums may be more open to the idea of sharing the photos on WikiArtifact, such as Museo Egizio. Contacting such institutions should be prioritized in order to get WikiArtifact started. Once there is a critical mass of participants, other institutions may be more motivated to join and permit photo distribution.

In addition to recording these permission rights in the Excel database, it is best practice to also make it clear to external users how these photos can be used. Creative Commons and Rightsstatements.org both offer simple, standardized language regarding reuse rights that will help external users of WikiArtifact understand how they can use these coffin images.[33][34] Dr. Cooney could review these licenses and statements, determine which best fits for each rights situation, and then describe or directly link to the appropriate license or rights statement.

b. Codex

[32] Cooney, Interview with Dr. Kara Cooney and her research team.
[33] "About The Licenses," Creative Commons, https://creativecommons.org/licenses/. (Accessed June 6, 2019).
[34] "Rights Statements," https://rightsstatements.org/page/1.0/?language=en. (Accessed June 6, 2019).

Dr. Cooney's data is qualitative, as it consists of Dr. Cooney's observations and determinations regarding signs of coffin reuse. Further, her research covers a very niche subject, with terms like "decorative reuse" and various reuse scores not necessarily ubiquitously used or known. Dr. Cooney's research would benefit greatly from a codex, which defines every data entry within the sheet. With regard to dates, for example, a codex would define what "19th Dynasty" means, or "mid." With regard to types of reuse, a codex would define "name reuse" and "decorative reuse," for example. The codex should be made available on the WikiArtifact website, to help users of the data as well as potential contributors to WikiArtifact, better understand the data and ensure consistent usage of terms.

B. Technical Rollout of WikiArtifact

a. Staffing limitations

While Dr. Cooney's research team does have some background with WordPress, it is possible that Dr. Cooney may need to hire a WordPress developer to create WikiArtifact. WordPress has a strong user community and offers robust technical documentation for setting up and running WordPress sites; however, if Dr. Cooney wants to make best use of advanced plug-ins, beyond what WordPress templates offer, she may need to hire a software engineer.

b. Collaborative workflow controls

Dr. Cooney wants WikiArtifact to be a collaborative site where researchers can contribute their findings and data. Dr. Cooney and her team will need to vet researchers to confirm both the veracity and style of the data contributed. For instance, Dr. Cooney may wish to ensure that external data does not exhibit problems with term variance, and enforce authority terms for certain topics. Depending on the volume of data received and the resources at Dr.

Cooney's disposal, she may not be able to do this on the largest scale, but she could narrow vocabulary control to a preferred term for specific topics, such as dating or names. Dr. Cooney and her team should consider implementing a vetting form, such as Google Forms or the WordPress API Gravity Forms.[35] These forms allow site administrators to control content that goes onto WikiArtifact before it is published on the site.

c. Maintaining the WordPress site and its plug-ins

WordPress is an open source platform, which allows programmers to create plug-ins to integrate into WordPress sites.[36] The WordPress platform has regular updates to "fix bugs and ensure speed, security, and compatibility."[37] Similarly, many programmers who have created WordPress plug-ins will update those APIs when WordPress runs a site-wide update.[38] However, if Dr. Cooney utilizes plug-ins that are not frequently updated by the developers who created them, she may need to consider hiring a programmer to fix the code. This would be an added cost, and it should be a consideration when integrating plug-ins not offered by preset templates.

V. Project Timeline

Dr. Cooney's timeline is greatly dependent on graduate student researcher availability and funding. We have created two project timelines that Dr. Cooney can follow at whatever pace her staffing and resources allow. We suggest Dr. Cooney first implement the data management recommendations (in turquoise), and then begin creating WikiArtifact (in pink).

Phase 1: Data Management

---

[35] "Using the API Lead Form with Gravity Forms in Wordpress," Tripleseat Support, accessed June 7, 2019.
[36] "Why Is WordPress Free? What Are the Costs? What Is the Catch?," WPBeginner, January 22, 2019, https://www.wpbeginner.com/beginners-guide/why-is-wordpress-free-what-are-the-costs-what-is-the-catch/.
[37] "Why Is WordPress Free? What Are the Costs? What Is the Catch?," WPBeginner.
[38] Ibid.

A. Understand the Schema: In order to implement the new schema that we have created, Dr. Cooney and her team should review the schema and its definitions to better understand where information falls in the database.

B. Finalize the Controlled Vocabulary: There were many variant terms that we were not able to create authority terms for because we did not have the Egyptology expertise to do so. We have shared these lists of variant terms with Dr. Cooney and her team. They will need to review these and come up with a controlled vocabulary they feel comfortable with.

C. Generate Unique Identifiers: Creating unique identifiers for each coffins will greatly facilitate data retrieval and long-term stewardship. Currently, data regarding each coffin is located in the Excel, photo files, and WikiArtifact. Generating unique identifiers for each coffin—and integrating these unique IDs into all three of these spaces—will allow Dr. Cooney and her team to uniquely, efficiently, and unambiguously identify each coffin, no matter where they are digitally stored.[39] Linking photos, Excel entries, and WikiArtifact in this way will also be extremely helpful for new graduate student researchers who are not familiar with the coffin reuse database. Additionally, these unique IDs can be incorporated into a preferred citation for coffins within WikiArtifact, so that external users of the data can easily locate and access the correct coffin.[40]

[39] "On the Utility of Identification Schemes for Digital Earth Science Data: An Assessment and Recommendations | SpringerLink," accessed June 8, 2019, https://link.springer.com/article/10.1007%2Fs12145-011-0083-6.
[40] Ibid.

Unique identifiers can be created through the Online UUID Generator, a site that generates a unique identifier using a timestamp and the MAC address of the computer to make it truly unique.[41] Dr. Cooney and her team may want to create unique identifiers that have visible characteristics tied to the coffin; for example, each unique ID could begin with the initials of the city where the museum is located.

D. Cleaning the Data: The data within the database must be placed in the new schema using the controlled vocabulary once the previous steps are completed. When we tested out implementing the schema, we found that transforming ten coffin entries takes around 30 minutes. Therefore, 300 coffin entries will take approximately 15 hours to input.

E. Create a Codex: This step will also be very time consuming. It is not the most time-sensitive item, as it will be most helpful for external users of WikiArtifact, in understanding what the data means. It is, however, an important element for creating the necessary context needed for resharing data.

Phase 2: WikiArtifact Implementation

A. Contact Museums for Permissions: Without permissions from the museums, the vision of WikiArtifact as a public site is not viable. Accordingly, this is a very important first step in setting up WikiArtifact.

[41] "Online UUID Generator Tool," https://www.uuidgenerator.net/. (Accessed June 7, 2019).

B. Update the Database with Museum Permissions: Once the museums are contacted, the Excel should be updated with the museum's response or terms for sharing so that this information is in a centralized location. Keeping the museums' responses organized will also allow Dr. Cooney to understand if her current vision of sharing coffin photos publicly is achievable in the format she wants.

C. Create a WordPress Site: During this phase, Dr. Cooney can determine what plug-ins are needed and if she would like to hire a programmer to facilitate this process. A good starting point would be to eliminate duplicate photos of coffins, to streamline storage and identify which photos should be featured on the website. Duplicate photo sorting software exists for both Windows and Mac computers.[42]

D. Batch Upload Excel and Photos: Excel files can be batch uploaded to WordPress using a plug-in.[43] Batch uploading will prevent Dr. Cooney's team from manually inputting data. Similarly, WordPress offers a plug-in that batch uploads media files, including photos.[44]

E. Implement Vetting Forms: A form needs to be created that vets researchers who will contribute to WikiArtifact. Dr. Cooney can determine which metadata fields are necessary for each contributor to include, and embed these into the form. She can also determine which fields should comply with a controlled vocabulary, and either provide for that via a dropdown menu where applicable or send documentation to contributors with the authority terms. Such a controlled form will give Dr. Cooney and her team time

---

[42] "Photos Duplicate Cleaner on the Mac App Store," accessed June 8, 2019, https://itunes.apple.com/us/app/photos-duplicate-cleaner/id592704001?mt=12;%20https://www.ashisoft.com/blog/top-5-best-duplicate-photo-finder-to-delete-duplicate-photos/.

[43] "Import Spreadsheets from Microsoft Excel – WordPress Plugin | WordPress.Org," accessed June 9, 2019, https://wordpress.org/plugins/import-spreadsheets-from-microsoft-excel/.

[44] "How to Bulk Upload WordPress Media Files Using FTP," WPBeginner, January 10, 2018, https://www.wpbeginner.com/plugins/how-to-bulk-upload-wordpress-media-files-using-ftp/.

to confirm the accuracy of contributed data before it goes live on the site, as well as introduce some standardization via the schema elements and controlled vocabulary.

VII. Conclusion

Dr. Cooney's research on coffin reuse in Ancient Egypt provides a fascinating look into the economies, social conditions, artistry, and spiritual beliefs of Ancient Egypt. Dr. Cooney's research has revealed some stunning discoveries—for instance, coffin reuse rates during the 21st averaged 60%, suggesting that the practice was socially acceptable and legal. Some coffins even reveal multiple reuses.[45] After nearly a decade of traveling the globe, Dr. Cooney's findings offer a rare and invaluable look at Ancient Egyptian society, culture, and beliefs.

Active data management will help ensure the usefulness and preservation of all of the valuable data Dr. Cooney has collected. In addition to facilitating preservation, the recommendations within this report will help foster structured data sharing and peer-to-peer collaboration. The new schema and controlled vocabulary are crucial steps to increase uniformity and consistency in the data management process.

The collaborative, visual nature of WikiArtifact has "the potential to revolutionize how we approach object studies in archaeology, art history, and Egyptology."[46] The use of WordPress will make WikiArtifact a space for collaboration, centralized data sharing, and equitable access to cultural heritage materials, while allowing Dr. Cooney to stay within her budget. Furthermore, Dr. Cooney's team's familiarity with WordPress will play an important role in the launch and longevity of the project. Ultimately, conscientious and proactive data management is critical to facilitating WikiArtifact's goals and long-term success.

---

[45] "Update from ARCE: Current Research, Excavation and Conservation Projects in Egypt," *NILE* Magazine, 60.
[46] Cooney, Wells, and Campbell, 2018, "National Geographic: Storytelling and Technology," 2.

VIII. References

"About The Licenses." n.d. Creative Commons. Accessed June 6, 2019. https://creativecommons.org/licenses/.

Borgman, Christine L. *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge: MIT Press, 2015.

Cooney, Kara. Interview with Dr. Kara Cooney and her research team. In-person, April 18, 2019.

Cooney, Kara, and Amber Wells. "Coffin List (FULL) 3.0," 2018.

Cooney, Kara, Amber Wells, and Rose Campbell. "National Geographic: Storytelling and Technology," 2018.

"DCMI: Dublin Core Metadata Element Set, Version 1.1: Reference Description." Accessed April 27, 2019. http://www.dublincore.org/specifications/dublin-core/dces/.

"DCMI Type Vocabulary." n.d. Dublin Core Metadata Initiative. Accessed June 7, 2019. http://www.dublincore.org/specifications/dublin-core/dcmi-type-vocabulary/.

"DCMI: Using Dublin Core." n.d. Accessed April 27, 2019. http://www.dublincore.org/specifications/dublin-core/usageguide/elements/.

"Getty Thesaurus of Geographic Names." n.d. Getty Research Institute. Accessed June 8, 2019. https://www.getty.edu/research/tools/vocabularies/tgn/.

"Getty Union List of Artist Names." n.d. Getty Research Institute. Accessed June 8, 2019. https://www.getty.edu/research/tools/vocabularies/ulan/.

"How to Bulk Upload WordPress Media Files Using FTP." WPBeginner, January 11, 2018. https://www.wpbeginner.com/plugins/how-to-bulk-upload-wordpress-media-files-using-ftp/.

"Import Spreadsheets from Microsoft Excel – WordPress Plugin | WordPress.Org." Accessed June 11, 2019. https://wordpress.org/plugins/import-spreadsheets-from-microsoft-excel/.

"MIDAS Heritage: The UK Historic Environment Data Standard." Forum on Information Standards in Heritage, October 2012. https://historicengland.org.uk/images-books/publications/midas-heritage/midas-heritage-2012-v1_1/.

Miller, Stephen J. 2011. *Metadata for Digital Collections: A How-to-Do-It Manual*. London, UK: Facet Publishing.

Niwinski, Andrzej. 1988. *21st Dynasty coffins from Thebes: chronological and typological studies*. Mainz am Rhein: P. von Zabern.

"Omeka Classic." n.d. Accessed June 6, 2019. https://omeka.org/classic/.

"Omeka.Net Price List." 2018. Corporation for Digital Scholarship.

"On the Utility of Identification Schemes for Digital Earth Science Data: An Assessment and Recommendations | SpringerLink." Accessed June 9, 2019. https://link.springer.com/article/10.1007%2Fs12145-011-0083-6.

"Online UUID Generator Tool." n.d. Accessed June 7, 2019. https://www.uuidgenerator.net/.

"Photos Duplicate Cleaner on the Mac App Store." Accessed June 8, 2019. https://itunes.apple.com/us/app/photos-duplicate-cleaner/id592704001?mt=12;%20https://www.ashisoft.com/blog/top-5-best-duplicate-photo-finder-to-delete-duplicate-photos/.

"Pricing." n.d. Omeka.Net. Accessed June 6, 2019. https://www.omeka.net/signup.

"Rights Statements." n.d. Accessed June 6, 2019. https://rightsstatements.org/page/1.0/?language=en.

Strasser, Carly. 2015. "Research Data Management." NISO Primer Series. Baltimore, MD: National Information Standards Organization.

"UCLA Library Events." n.d. UCLA Library. Accessed June 8, 2019. https://www.library.ucla.edu/events/data-cleaning-openrefine.

"Update from ARCE: Current Research, Excavation and Conservation Projects in Egypt." *NILE Magazine*, November 2018.

"Using the API Lead Form with Gravity Forms in Wordpress." Tripleseat Support. Accessed June 9, 2019. http://tripleseat.zendesk.com/hc/en-us/articles/219006788-Using-the-API-Lead-Form-with-Gravity-Forms-in-Wordpress.

"What Is CPanel? How to Use CPanel for WordPress Hosting." WPBeginner. Accessed June 7, 2019. https://www.wpbeginner.com/glossary/cpanel/.

"Why Is WordPress Free? What Are the Costs? What Is the Catch?" WPBeginner, January 22, 2019. https://www.wpbeginner.com/beginners-guide/why-is-wordpress-free-what-are-the-costs-what-is-the-catch/.

Wolf, Damon. Interview with Technical Account Manager | Information Technology Services at the University of California, Los Angeles. Phone, April 17, 2019.

Woodley, Mary S. 2016. "Setting the Stage." In *Introduction to Metadata*, edited by Murtha Baca. http://www.getty.edu/publications/intrometadata/metadata-matters/.

"WordPress.Com Plans and Pricing – Get Started for Free Today!" *WordPress.Com* (blog), February 23, 2016. https://wordpress.com/pricing/.

IX. Appendix

A. Revised Metadata Schema Definitions

The following are definitions of each element of the revised metadata schema. Understanding these will facilitate consistent data entry. We have starred (*) elements that were carried forward from Dr. Cooney's initial metadata schema. Other elements that are not starred represent revisions and additions we made.

Provenance metadata

- **City of holding institution***: City in which the institution that holds the coffin is located. Best practice is to use the Getty Thesaurus of Geographic Names (TGN) vocabulary.[47]
- **Holding institution***: Institution that holds the coffin. Best practice is to use the Getty Union List of Artist Names (ULAN) vocabulary.[48]
- **Accession number***: The holding institution's unique identifier for the coffin.
- **Niwinski number***: Correlating coffin number in the Niwinski study.[49]
- **Acquisition date**: Date in which the coffin was acquired by the museum. Best practice is to use the format MM-DD-YYYY.
- **Purchase location**: Location in which the coffin was acquired by the museum. Best practice is to use the TGN vocabulary.
- **Seller**: Agent who sold the coffin to the holding institution. Best practice is to use the ULAN vocabulary.
- **Buyer**: Agent who acquired the coffin for the holding institution.. Best practice is to use the ULAN vocabulary.
- **Current collection**: Collection in which the coffin is currently housed in within the holding institution.
- **Excavation location**: Location in which the coffin was excavated. Best practice is to use the TGN vocabulary.
- **Excavation date**: Date(s) in which the coffin was excavated. Dr. Cooney can make a decision as to whether the date range or final date of excavation is preferred. Best practice is to use the format MM-DD-YYYY–MM-DD-YYYY.
- **Excavation team/agent**: Team or agent that excavated the coffin. Best practice is to use the ULAN vocabulary.
- **Date examined***: Date in which the coffin was examined for the purposes of this coffin reuse study. Best practice is to use the format MM-DD-YYYY.

---

[47] "Getty Thesaurus of Geographic Names," Getty Research Institute, https://www.getty.edu/research/tools/vocabularies/tgn/. (Accessed June 8, 2019).
[48] "Getty Union List of Artist Names," Getty Research Institute, https://www.getty.edu/research/tools/vocabularies/ulan/. (Accessed June 8, 2019).
[49] Andrzej Niwinski, 1988, *21st Dynasty coffins from Thebes: chronological and typological studies*, Mainz am Rhein: P. von Zabern.

- **Data collector**: Agent who collected qualitative data on the coffin as part of this coffin reuse study.
- **Reuse observations and explanation**: Observations about the coffin's reuse, and justification for why the coffin received its reuse score.

Descriptive metadata

- **Coffin type***: When we queried Dr. Cooney's team for this definition, we received the following: "if it's inner or outer of board or mask." This is an example of "definition by example," and is not best practice because defining something by listing examples of it does "not establish clear boundaries between what is and is not included in a concept."[50] We recommend that Dr. Cooney offer a more robust definition to facilitate better understanding of this metadata field.
- **Ambiguity and/or notes on Coffin Type**: Any additional notes on the coffin type that go beyond defining the actual coffin type.
- **Coffin part***: When we queried Dr. Cooney's team for this definition, we received the following: "is if it's only lid or case or a fragment." Again, this is "definition by example," and is not best practice. We recommend that Dr. Cooney offer a more robust definition to facilitate better understanding of this metadata field.
- **Ambiguity and/or notes on Coffin Part**: Any additional notes on the coffin type that go beyond defining the actual coffin type.
- **Coffin time period descriptor (early, mid, late)**: A general descriptor for the time period of the coffin as being either "early," "mid," or "late." Dr. Cooney and her team may wish to define these more specifically, either here or in the project's codex (by number of decades, for example, represented in each stage).
- **Coffin start time period**: The starting Dynasty in which the coffin could be dated.
- **Coffin end time period**: The ending Dynasty in which the coffin could be dated.
- **Coffin date ambiguity**: A field to denote any ambiguity or uncertainty about the dating. Use the term "Yes" to denote that the period is uncertain and "No" to denote that it is not.
- **Additional dates**: Any other dates associated with the coffin and its items. For example, if mummy linens placed on the body found inside the coffin are inscribed with dates that differ from the coffin date (ie. the date of the interment of the body differs from the date of the coffin).
- **Name(s) of the deceased***: Names of deceased person(s) who have used the coffin.
- **Title***: Where applicable, the coffins reflect the museum's title for the object. Sometimes the title is chosen by Dr. Cooney and her research team. An explanation here of when and why Dr. Cooney sometimes chooses a new title would be helpful.

[50] Christine L. Borgman, *Big Data, Little Data, No Data: Scholarship in the Networked World* (Cambridge: MIT Press, 2015), 19.

- **Reuse score***: A rating of Dr. Cooney's confidence in her ability to see coffin reuse on a scale from 0 to 3, with 3 being obvious and clearly visible evidence of reuse, 1 being only circumstantial, and 0 being no visible evidence of reuse. To clarify, a 0 score does not mean that a given coffin was not reused; it just means that Dr. Cooney cannot see evidence of that (evidence of reuse could be carefully covered by a carpenter, for example).
- **Type(s) of reuse***: Dr. Cooney's determination of the types of reuse in the coffin. Terminology should come from a controlled list of terms determined by Dr. Cooney.
- **Relation**: Any notes regarding the coffin's relation to other coffins on the list. Include the unique identifier of the other coffin in this field as well.

Administrative metadata

- **Unique identifier**: Unique identifier for each coffin generated by Dr. Cooney's team.
- **Rightsholder**: The institution or agent with which Dr. Cooney and her team correspond concerning access rights for the coffin photos and data.
- **Access rights**: Information from museum permissions concerning who is allowed to see coffin photos and data.
- **File location notes**: Notes about locations of photos and field note files, including when these files are missing.

B. Dublin Core Crosswalk

The following is a crosswalk that maps Dr. Cooney's revised metadata schema onto Dublin Core.

- DC: Title
    - KC: Title
- DC: Creator
    - Not listed on Dr. Cooney's current spreadsheet. It is recommended that this field be populated the "Unknown," as the creators of these coffins are not known.
- DC: Subject
    - KC: Types of reuse
    - KC: Name of the deceased
- DC: Description
    - KC: Coffin type
    - KC: Coffin part
    - KC: Name of the deceased
    - KC: Reuse score
    - KC: Acquisition date
    - KC: Purchase location
    - KC: Seller
    - KC: Buyer
    - KC: Current collection
    - KC: Excavation location
    - KC: Excavation date
    - KC: Excavation team/agent
- DC: Publisher
    - KC: Holding institution
- DC: Contributor
    - KC: Data collector
- DC: Date
    - KC: Coffin time period descriptor (early, mid, late)
    - KC: Coffin start time Period
    - KC: Coffin end time period
    - KC: Coffin date ambiguity
    - KC: Acquisition date
    - KC: Excavation date
    - KC: Date examined
- DC: Type

    - Not listed on Dr. Cooney's current spreadsheet. It is recommended that this field be populated with both PhysicalObject and Dataset. Both of these terms are taken from the recommended DCMI Type vocabulary.[51]
- DC: Identifier
    - KC: Unique identifier
- DC: Language
    - Not listed on Dr. Cooney's current spreadsheet. Where there is writing on the coffin, it is recommended that Egyptian is cited using the recommended standards for Dublin Core: RFC 3066 and ISO 39, which define primary language tags and subtags.[52]
- DC: Relation
    - KC: Relation
- DC: Coverage
    - KC: Coffin time period descriptor (early, mid, late)
    - KC: Coffin start time Period
    - KC: Coffin end time period
    - KC: Coffin date ambiguity
    - KC: Excavation location
- DC: Rights
    - KC: Access rights
    - KC: Rightsholder

[51] "DCMI Type Vocabulary," Dublin Core Metadata Initiative, http://www.dublincore.org/specifications/dublin-core/dcmi-type-vocabulary/. (Accessed June 7, 2019).
[52] "DCMI: Dublin Core Metadata Element Set, Version 1.1: Reference Description," http://www.dublincore.org/specifications/dublin-core/dces/. (Accessed April 27, 2019).

C. Controlled Vocabulary Guidelines & Examples
In addition to eliminating variant terms via OpenRefine, we implemented the following general guidelines for standardizing and controlling the vocabulary throughout the spreadsheet.

Style Conventions:

- Use sentence case throughout
- Use commas to denote multiple items (not plus signs or ampersands)
- For specific dates or date ranges in which the month, day, and/or year are known, use MM-DD-YYYY or MM-DD-YYYY–MM-DD-YYYY.
- For the "Coffin Date Ambiguity" field, enter either Yes or No
- For the "Coffin time period descriptor" field, enter either "early," "mid," "late," or "N/A"
- Do not leave any column blank. For data that was not recorded, use "N/R"

External Thesauri and Controlled Vocabularies:
The following thesauri should be used in the denoted fields where applicable. It is marked when use of these controlled vocabularies is required; otherwise, it is only considered best practice.

The Getty Thesaurus of Geographic Names:

- City of Holding Institution
- Purchase Location
- Excavation Location

The Getty Union List of Artist Names:

- Holding Institution
- Seller
- Buyer
- Excavation Team/Agent

When mapping elements onto Dublin Core, the following vocabularies should be used:

- DCMI Type Vocabulary:[53] Type (required)
- RFC 3066 and ISO 39:[54] Language

[53]"DCMI: Dublin Core Metadata Element Set, Version 1.1: Reference Description," http://www.dublincore.org/specifications/dublin-core/dces/. (Accessed April 27, 2019).
[54] "DCMI: Using Dublin Core," http://www.dublincore.org/specifications/dublin-core/usageguide/elements/.

The following is an example of controlled vocabulary work through OpenRefine, taking these stylistic guidelines into account.

The unbulleted terms are the authority terms that should be used in place of all of the variant terms, which are bulleted underneath that term. Items marked with an asterisk (*) denote data entries with extra details that would be placed in the newly created "Ambiguity and/or Notes on Coffin Type" column.

Dr. Cooney's team will need to review these recommendations, given their Egyptology expertise.

COFFIN TYPE

N/R

- Blank

Inner coffin

- Inner Coffin
- inner coffin
- Inner coffin (seems to be Stola)*

Inner coffin, mummy board

- Inner coffin + mummy board
- Inner coffin/Mummy board

Outer coffin

- Outer Coffin
- Outer coffin (?)*
- Outer(?) coffin*

Outer coffin, inner coffin

- Outer coffin + inner coffin

Outer coffin, inner coffin, mummy board

- Outer coffin + inner coffin + mummy board

Stola coffin, inner coffin

- Stola, Inner coffin

D. Example Set of Cleaned Data

PROVENANCE METADATA														
City of Holding Institution	Holding Institution	Accession Number	Niwinski Number	Acquisition Date	Purchase Location	Seller	Buyer	Current Collection	Excavation Location	Excavation Date	Excavation Team/Agent	Date Examined	Data Collector	Reuse Observations and Explanation
Turin, Italy	Museo Egizio di Torino	2212; CG 10112-10114	377	1822	N/R	B. Drovetti	N/R	N/R	N/R	N/R	N/R	06-07-2011	Kara Cooney	Decorative reuse. Very thin plaster layer. No traces of plaster for reuse. Very nice wood. One thick plank for mummy board. Coffin lid base is also one thick plank, but more boards used. Case bottom of coffin uses many pieces of local wood, but not important for viewer to see quality. Sides of case, though are thick, high quality wood. Different woods suggestive of reuse from older coffins. So, probably the plaster layer is so thin because the wood is so nice. This is a nice version of simple dot decoration, but no originality. Case sides very repetitive and generic. Name on the lid original in polychrome. Title also in polychrome.
Turin, Italy	Museo Egizio di Torino	2212; CG 10112-10114	377	1822	N/R	B. Drovetti	N/R	N/R	N/R	N/R	N/R	06-07-2011	Kara Cooney	N/R
Turin, Italy	Museo Egizio di Torino	2214; CG 10121	None	N/R	N/R	N/R	N/R	N/R	N/R	N/R	N/R	N/R	Kara Cooney	N/R
Turin, Italy	Museo Egizio di Torino	2217; CG 10110	378	1822	N/R	B. Drovetti	N/R	N/R	N/R	N/R	N/R	06-07-2011	Kara Cooney	Just coffin case with no lid. Woman's coffin with no signs of reuse. Coffin wood insanely nice. Sides are each complete boards. Smaller pieces by head but just to create the round shape. No raised relief though. Thinner layer of plaster and no plaster underneath with any previous decoration. Nice would could be a marker that the coffin was reused from temple wood or other coffin wood. Similar in decoration to Hennatawy in Boston, but not raised relief. Most interesting things for defensive burial - no gilding in combination with really nice wood, probably imported. Probably orpiment and yellow ochre.
Turin, Italy	Museo Egizio di Torino	2219; CG 10117a-b	379	1822	N/R	B. Drovetti	N/R	N/R	N/R	N/R	N/R	06-07-2011	Kara Cooney	Decorative reuse, Blank space for name, gender modification. Breasts removed. Name space erased but part of name of title is still underneath. This is interesting and suggests reuse. Title is wsir nbt pr Smayt _____ r nb n ist mwt nTr. Stnage. Why erase part of the name? Is it part of a title? Then, why erase it? Nothing seems ever written...sparkly yellow bits in paint so maybe orpiment. But, there is yellow paint! And, on second glance, varnish! So, whole coffin has orpiment/yellow ochre. More ochre and thus mustard color to coffin. Not best craftsmanship, but blu pigment and orpiment (so type II coffin).
Turin, Italy	Museo Egizio di Torino	2221; CG 10116a-b	380	1822	N/R	B. Drovetti	N/R	N/R	N/R	N/R	N/R	06-07-2011	Kara Cooney	Decorative reuse, gender modification. Male coffin changed to woman's, beard hole covered. Usurpation only clear because of the beard hole under the chin! Cracking around the lappets of the wig suggests addition of plaster for headcloth. Text at feet ends with Wsir only so maybe they never added the name in. Raised plaster relief of sundisks and figures. Yellow paint. No breasts added to this usurpation by a woman of a man's coffin. With this restoration, if there was a 19th Dynasty coffin underneath, I wouldn't see it.
Turin, Italy	Museo Egizio di Torino	2222; CG 10111a-b	381	1822	N/R	B. Drovetti	N/R	N/R	N/R	N/R	N/R	06-07-2011	Kara Cooney	No lid to outer coffin. Not much interest in this piece as I remember.
Turin, Italy	Museo Egizio di Torino	2226; CG 10104a-b	382	1822	N/R	B. Drovetti	N/R	N/R	N/R	N/R	N/R	06-07-2011	Kara Cooney	The painting on the case side interior is unfinished! White line on red background. Masculine Issi on case right side! God with label of Isis. Pyramid with tree goddess is breast shaped while pyramid with Hathor of western mountain is sharp. This coffin is most interesting for male-female differences. For instance, strange that a male deceased is getting water from the tree goddess even though the coffin was made for the female Tabakenkhonsu. Female deceased is depicted once on each side and is carefully labeled.
Turin, Italy	Museo Egizio di Torino	2226; 10105	382	1822	N/R	B. Drovetti	N/R	N/R	N/R	N/R	N/R	06-07-2011	Kara Cooney	Mummy board has same style of raised relief (as 2226; CG 10104a-b inner coffin). If compare piece to piece, the inner coffin is higher in value than mummy board.
Turin, Italy	Museo Egizio di Torino	2227; CG 10115	383	1822	N/R	B. Drovetti	N/R	N/R	N/R	N/R	N/R	06-07-2011	Kara Cooney	Deceased in front of a number of dvinities. Nothing usurped is visble to me. Decent wood.
Turin, Italy	Museo Egizio di Torino	2228; CG 10119a-b	384	1822	N/R	B. Drovetti	N/R	N/R	N/R	N/R	N/R	06-06-2011	Kara Cooney	Multiple reuse, decorative reuse, markers of Ramesside, blank space for name. Whole surface of the lids is covered with a layer of linen and then plaster so no indication of white dress with red pleasts that must have been underneath. Dyn 19 according to feet and female shape of body. Label says "daily dress"... Sandals added to the feet in usurpation! The only instance of the dead's name is on the strip between the feet on the lid. Case sides--text space for naming female deceased is left out. Case right side--with tree goddess scene called nbt pr Smayt. In large format next to small caption we read Wsir nbt pr Smayt n imn but blank where name should be! Case left side in polychrome Wsir nbt pr Smayt n imn and erased name. But a mn sign is visible! So, maybe this is twice usurped.
Turin, Italy	Museo Egizio di Torino	2228; CG 10120	384	1822	N/R	B. Drovetti	N/R	N/R	N/R	N/R	N/R	06-06-2011	Kara Cooney	No text on mummy board at all!
Turin, Italy	Museo Egizio di Torino	2236-2237; CG 10101a-b	385	1822–1824	N/R	B. Drovetti	Museo Egizio di Torino	N/R	probably from TT 291	1818	B. Drovetti	06-05-2011	Kara Cooney	Decorative reuse. Painted plaster visible under surface on right shoulder; shoulders shaved down on inside to fit inner coffin, so not made as set. The traces of reuse are few, but one detail of painted plaster under coffin's right shoulder. Got good detail photographs. Plaster on lid much thicker than plaster on case. But, on both, the plaster is rough and lumpy, as if an older coffin was choseled down of its decoration, but didn't plane the wood down before they did the decoration.
Turin, Italy	Museo Egizio di Torino	2236-2237; CG 10102a-b	385	1822–1824	N/R	B. Drovetti	Museo Egizio di Torino	N/R	probably from TT 291	1818	B. Drovetti	06-05-2011	Kara Cooney	Decorative reuse. Coffin very small and shaved down on the inner shoulders, probably because the mummy didn't fit into the coffin with the lid. The inner coffin and outer coffin are way out of sync for a set. These can't have been made for one another but were pieced together. The inner and outer coffin also do not fit in terms of decoration. The outer has large sections of unvarnished white. The inner is completely varnished a dark amber. Outer coffin has the Nut - Geb scene in same place as teh inner coffin, but outer coffin has ithyphallic Geb.
Turin, Italy	Museo Egizio di Torino	2236-2237; CG 10103	385	1822–1824	N/R	B. Drovetti	Museo Egizio di Torino	N/R	probably from TT 291	1818	B. Drovetti	06-05-2011	Kara Cooney	N/R
Turin, Italy	Museo Egizio di Torino	2238; CG 10106a-b, 10107	386	1822	N/R	B. Drovetti	N/R	N/R	N/R	N/R	N/R	06-05-2011	Kara Cooney	Decorative reuse, name reuse. Thick smooth grained boards of wood. If not imported wood, then a finer native variety. Fair to say that 21st Dynasty wood is an improvement over most 19th Dynasty wood. The inside of the coffin is undecorated except for the floor, which has a goddess. That and the arms on the coffin lid suggest an early 21st Dynasty date, but N says mid-21st D. Nice example of simple dot version decoration. So, wood and decoration quality do not meet. Name of the case left side next to worshipping deceased added later in black paint under polychrome Wsir. Text in really bad shape and illegible to me, but they read Knsw-ms according to the label. But name of original deceased in written. Deceased on coffin case right side with arms upraised is called Ba-pw-n. Also, there is a small kneeling figure in the scale scene on case right with same name Ba-pw-n. Not in Ranke.
Turin, Italy	Museo Egizio di Torino	2238; CG 10106a-b, 10107	386	1822	N/R	B. Drovetti	N/R	N/R	N/R	N/R	N/R	06-05-2011	Kara Cooney	Decorative reuse, name reuse. Mummy board is one piece of wood like Hori. In fact, decoration is similar, too. Maybe related? Board has name reinscribed in black, too. Top starts with Dd mde in Wsir Bapwn Dd.f ... and then invocation to Nut for wings and imperishable ones.
Turin, Italy	Museo Egizio di Torino	CG 10109	387	N/R	N/R	N/R	N/R	N/R	Dayr al-Madinah	1905	Schiaparelli	06-08-2011	Kara Cooney	N/R
Turin, Italy	Museo Egizio di Torino	Suppl. 7715; CG 10108a-b	388	N/R	N/R	N/R	N/R	N/R	Dayr al-Madinah	1905	Schiaparelli	06-08-2011	Kara Cooney	Decorative reuse, gender modification. Man's coffin usurped by a woman. Visible because of the beard hole. Yellow paint. Varnish not evident. Male deceased depicted but not female. Usurpation of head cloth very interesting. Hard to see what they did but the face is poorly formed. Was it re-carved? Interior of coffin case is monochrome. Only red, black, and yellow, but underside of the lid has image of Osiris in polychrome. Really nice. Four sons of Horus as canopic jars. Pomegranates under offering table of one. Gorgeous solar scene of arms and tree and ba on case's left side. Very interesting coffin.
Turin, Italy	Museo Egizio di Torino	2226	383	N/R	N/R	N/R	N/R	N/R	N/R	N/R	N/R	N/R	Kara Cooney	N/R

The following fields need to be standardized by Dr. Cooney's team, given their Egyptology expertise: title, name(s) of the deceased, types of reuse.
The following fields willbe completed by Dr. Cooney's team once they implement our plan: rightsholder, access rights, unique identifier.

DESCRIPTIVE METADATA									
Coffin Type	Ambiguity and/or notes on Coffin Type	Coffin Part	Ambiguity and/or notes on Coffin Part	Coffin Time Period descriptor (early, mid, late)	Coffin Start Time Period	Coffin Time Period descriptor (early, mid, late)	Coffin End Time Period	Coffin Date Ambiguity	Additional Dates
Inner coffin	N/R	Case, lid	N/R	mid	21st Dynasty	mid	21st Dynasty	No	N/R
Mummy board	N/R	N/R	N/R	mid	21st Dynasty	mid	21st Dynasty	No	N/R
Coffin	N/R	N/R	N/R	N/R	N/R	N/R	N/R	Yes	N/R
Outer coffin	N/R	Case	N/R	mid	21st Dynasty	mid	21st Dynasty	No	N/R
Inner coffin	N/R	Lid	This piece also has a mummyboard according to Notability file. The mummy board has a Blank Space for Name.	late	21st Dynasty	late	21st Dynasty	No	N/R
Inner coffin	N/R	Case, lid	N/R	late	21st Dynasty	late	21st Dynasty	No	N/R
Inner coffin	N/R	Case, lid	N/R	mid	21st Dynasty	late	21st Dynasty	No	N/R
Inner coffin	N/R	Case, lid	N/R	mid	21st Dynasty	mid	21st Dynasty	No	N/R
Mummy board	N/R	N/R	N/R	mid	21st Dynasty	mid	21st Dynasty	No	N/R
Outer coffin	N/R	Case	N/R	mid	21st Dynasty	late	21st Dynasty	No	N/R
Inner coffin	N/R	Case, lid	N/R	late	21st Dynasty	late	21st Dynasty	No	N/R
Mummy board	N/R	N/R	N/R	late	21st Dynasty	late	21st Dynasty	No	N/R
Outer coffin	N/R	Case, lid	N/R	mid	21st Dynasty	mid	21st Dynasty	No	N/R
Inner coffin	N/R	N/R	N/R	mid	21st Dynasty	mid	21st Dynasty	No	N/R
Mummy board	N/R	N/R	N/R	mid	21st Dynasty	mid	21st Dynasty	No	N/R
Inner coffin	N/R	Case, lid	N/R	mid	21st Dynasty	mid	21st Dynasty	No	N/R
Mummy board	N/R	N/R	N/R	mid	21st Dynasty	mid	21st Dynasty	No	N/R
Mummy board	N/R	N/R	N/R	mid	21st Dynasty	mid	21st Dynasty	No	N/R
Inner coffin	N/R	N/R	N/R	mid	21st Dynasty	mid	21st Dynasty	No	N/R
Outer coffin	N/R	Case	N/R	N/R	N/R	N/R	N/R	Yes	N/R

					ADMINISTRATIVE METADATA			
Name(s) of the Deceased	**Title**	**Reuse Score**	**Type(s) of Reuse**	**Relation**	**Unique Identifier**	**Rightsholder**	**Access Rights**	**File Location Notes**
N/R	N/R	1	Decorative Reuse; Mismatched Ledges;Mismatched Construction & Decoration	Coffin #277	N/R	N/R	N/R	N/R
N/R	N/R	1	N/R	Coffin #277	N/R	N/R	N/R	N/R
Anonymous	N/R	TBD	N/R	N/R	N/R	N/R	N/R	Notability file contains no notes.
Nesykhonsu	N/R	1	Mismatched Construction & Decoration	N/R	N/R	N/R	N/R	N/R
Anonymous	anonymous; Wsir nbt pr Smayt ___ ___f nb ist mwt nTr (?)	2	Decorative Reuse; Blank Space for Name; Gender Modification; Blank Space for Title; Plaster Modification	Coffin #379	N/R	N/R	N/R	N/R
anonymous man usurped by a woman	anonymous man usurped by a woman	2	Decorative Reuse; Gender Modification; Plaster Modification	Coffin #380	N/R	N/R	N/R	N/R
Bakenkhonsu	N/R	3	Plaster Modification; Wood Modification; Mismatched Ledges	Belongs in a set with Turin 2227 (Niwinski 383). Coffin #382	N/R	N/R	N/R	N/R
Tabakenkhonsu	N/R	0	No visible reuse	Coffin #382	N/R	N/R	N/R	N/R
Tabakenkhonsu	N/R	0	No visible reuse	Coffin #382	N/R	N/R	N/R	N/R
Tabakenkhonsu	N/R	0	No visible reuse	Coffin #381 A little confused with this. C.2226 is the inner and outer coffin. A few lines down, you have what appears to be another line for the same outer coffin.	N/R	N/R	N/R	N/R
chantress of Amun	N/R	3	Multiple Reuse; Decorative Reuse; Markers of Ramesside; Blank Spack for Name	Coffin #384	N/R	N/R	N/R	N/R
chantress of Amun	N/R	0	No visible reuse	Coffin #384	N/R	N/R	N/R	N/R
Butehamun	N/R	3	Decorative Reuse; Wood Modification; Plaster Modification; Mismatched Lid & Case; Mismatched Ledges	Coffin #385	N/R	N/R	N/R	N/R
Butehamun	N/R	2	Decorative Reuse; Wood Modification; Plaster Modification; Mismatched Lid & Case; Mismatched Ledges; Contextual Reuse	Coffin #385 This is another scenerior where the pices between inner, outer, and mummyboard are stylistically mismatched, and it is only provanance that tells us they go together. I don't know what type of reuse this is.	N/R	N/R	N/R	N/R
Butehamun	N/R	2	Markers of Ramesside	Coffin #385	N/R	N/R	N/R	N/R
Khonsumes, reused by Bapu	N/R	3	Decorative Reuse; Name Reuse; Mismatched Construction & Decoration; Mismatched Ledges	Coffin #386	N/R	N/R	N/R	N/R
Khonsumes, reused by Bapu	N/R	3	Decorative Reuse; Name Reuse	Coffin #386	N/R	N/R	N/R	N/R
Mwt-n-pr-imn	N/R	0	No visible reuse	Coffin #387	N/R	N/R	N/R	N/R
Herpeniset	N/R	1	Decorative Reuse; Gender Modification; Mismatched Leges	Coffin #388	N/R	N/R	N/R	N/R
N/R	N/R	0	Nothing usurped is visible	Coffin #383. Deceased in front of a number of divinities. Decent wood.	N/R	N/R	N/R	N/R

Core Coursework

Systems and Infrastructures // IS 270
Jean-François Blanchette
Winter 2019

Abstract
The internet has undeniably shaped the nature of contemporary research and information seeking, as more and more people turn to Google as their first line of inquiry, for better or for worse. Digital collections management systems provide a way to bring special collections online, facilitating increased access and scholarship. Addressed to the UCLA Library as a potential buyer of digital collections management software, this policy brief provides an assessment of digital collections management systems, complete with a description of the technology, current issues, and future trends.

Inclusion in portfolio
This is an important addition to my portfolio in terms of both content and style. With regard to content, digital collections management systems are relevant to my current work at the Digital Library Program, and could very much be a part of my future employment. It was deeply beneficial for me to investigate the software on a number of fronts, from architecture and interoperability to standards and markets. With regard to style, this assignment required synthesis and precision, as it distilled 6,000 words of writing completed over the course of six weeks into 2,000 words. This writing style felt distinct from much of the academic writing completed during my time in the program, and especially relevant for my professional development.

March 18, 2019

Ms. Virginia Steele
Norman and Armena Powell University Librarian
UCLA Library
280 Charles E Young Dr N
Los Angeles, CA 90095

Re: Digital Collections Management Systems

Dear Ms. Steele,

Please find enclosed an assessment of digital collections management systems, complete with a description of the technology, current issues it is facing, and future trends. The internet has undeniably shaped the nature of contemporary research and information seeking, as more and more people turn to Google as their first line of inquiry, for better or for worse. Digital collections management systems provide a way to bring UCLA Library's robust special collections online, facilitating increased access and scholarship.

That said, there are some facets of the technology that should be carefully evaluated before investing considerable resources into building out UCLA's digital collections. The technology's high switching costs and propensity for lock-in means that a very thoughtful decision should be made at the outset as to which platform to use. The technology also opens institutions to increased liability regarding copyright protection, given the exposure an online audience entails. Other considerations regarding the software's technical architecture and web standards will inevitably affect the technology's development in the long term. These concerns, however, do not necessarily outweigh the technology's potential benefits to the UC scholarly community, but instead should be carefully appraised and considered when selecting a software solution.

Thank you for your time in considering this technology for UCLA Library. Please feel free to be in touch directly should you have any questions or concerns.

Yours sincerely,

Savannah Lake
Department of Information Studies, UCLA

Digital Collections Management Systems

Overview

- Digital collections management systems integrate cataloging databases with web platforms, allowing collecting institutions to bring their collections online.
- While the market is diffuse and customers have a wide variety of platforms to choose from, the software exhibits high switching costs and lock-in, and is not built to be interoperable with other digital collections.
- The content-rich technology already faces challenges on slower internet connections, which will likely worsen as the internet at large confronts capacity problems in the coming years.
- Institutions expose themselves to greater risk regarding copyright infringement with the technology, and also must be aware of how web standards created by largely corporate interests may disadvantage disenfranchised or disabled users.
- The design of the systems continues to evolve, with improvements in user experience and technical capacity. However, this progress could be impeded by the economic instability of the technology's major client base.

Introduction

Digital collections management systems allow cultural heritage institutions to bring collections online, reaching patrons near and far. This dynamic technology encompasses multiple computing functions, including online networking, extensive file storage (often of various formats), and database cataloging. This briefing examines attributes of the technology, its challenges, and future trends.

Background

As libraries, archives, and museums confront the digital age of information and strive to stay relevant, more cultural heritage institutions are taking their collections online through digital collections management systems. These technologies are highly complex, integrating cataloging databases and records management systems with web platforms.

Customers have a wide variety of digital collections management systems to choose from. They can build their own customized solution, utilizing applications like Fedora and Ruby on Rails to build out databases and websites. If they already have a collections management system in place, like PastPerfect or Axiell, they can often buy the digital collections component as an add-on service. Or they can choose a system that focuses entirely on digital collections, like CONTENTdm. Some of these options are proprietary, with strong customer-service support and corresponding fees, while others are free and open-source platforms, requiring instead that the user have substantial technical know-how. Accordingly, the product options and market structure are rather diffuse, allowing institutions to choose solutions that best fit the needs of their collection and staff.

This multitude of options, however, does not necessarily enable customer mobility. While the market does encourage competition on functionality and price, high switching costs and lock-in are prominent features of digital collections management systems—making selecting the correct platform at the outset a high priority.

Technology Description

Collections management solutions allow users to organize, search, and manage their collections. Through the software, users can ascribe metadata to their holdings within a database of records. Digital collections management systems build on these solutions, making collections available online.

These technologies can take the form of installed software or web applications. Both forms represent a fundamentally different approach to computing. Web-based solutions follow the client-server architectural model, with processing and storage performed on the server side, in the cloud. Popular solutions that follow this model include CONTENTdm. Installed software packages, on the other hand, live in computers or on a shared internal server, with processing and storage taking place locally. Popular solutions include PastPerfect and Axiell. Some of these solutions, including PastPerfect, will offer to host customer files online in cloud storage, at a fee.[1]

[1] "PastPerfect Online: Expand Your Audience by Providing Online Access," http://www.museumsoftware.com/pponline.html, (February 18, 2019)

Competitive digital collections management solutions are compatible with any browser and on any device, including laptops, tablets, smart phones, and PC and Mac systems. In addition to interoperability with web browsers and devices, digital collections management solutions can offer application programming interfaces (APIs) to more deeply embed the collection with other web functions. CONTENTdm, for example, developed its own API, which allows institutions to customize their collections with visual branding, maps, Drupal, and online shopping carts.[2]

Digital collections management software is largely used in professional settings by cultural heritage institutions. As such, regulations inherent to the profession influence the development of the technology—in particular, the use of metadata best practices. Digital collections management systems are designed with metadata input fields, which allow institutions to ascribe administrative, technical, and descriptive metadata to an object. Some digital collections management systems even integrate known metadata standards like Dublin Core and DACS (Describing Archives: A Content Standard) or controlled vocabularies (in the form of thesauri and authority lists) directly into the software.[3]

Key Challenges and Issues

Digital collections management systems confront a number of challenges, spanning the expanse of their technical architecture, design, interoperability, regulation, and standards.

1. High lock-in and switching costs

Digital collections management software—particularly the packaged solutions, like CONTENTdm, Axiell, and PastPerfect—exhibit lock-in and high switching costs. At the outset, institutions that already have a collections management solution in place, such as Axiell or PastPerfect, likely feel locked into that solution, finding it most convenient to simply extend the existing software to add on the digital collections component instead of using a different digital software solution.

This lock-in is compounded by the high switching costs that all of the solutions exhibit—even the more modular, homegrown solutions customers build themselves through applications like Fedora and Ruby on Rails. While most digital collections management software will

[2] "CONTENTdm Features," https://www.oclc.org/en/contentdm/features.html, (February 2, 2019)
[3] "Resources," https://www.oclc.org/en/contentdm/resources.html, (February 10, 2019)

facilitate switching from one software solution to another by providing XML or DBF file exports, switching still requires time and resources, chiefly in standardizing data for migrations.[4] This means that there are high costs to switching platforms, even for open source solutions.[5]

Further, digital collections are not designed to interoperate with one another. This siloed approach forces users to look for items within each institution's digital collection instead of searching one universal collection that includes numerous institutions. Currently, the process of combining different digital collections is tedious, involving harvesting and standardizing metadata records, creating a website for the merged collections, and developing an API that allows the collections to be shared with external sites.[6] Some data aggregation tools can perform these functions, but the effort still requires time, staff, and resources. This has implications for UCLA libraries, if there is any desire to coordinate with the UC system at large to support more robust searching. The logistics of UC-wide coordination are especially challenging if certain UC institutions are already locked in to a certain platform.

2. Image-heavy technology requires sizeable bandwidth

Websites displaying digital collections are necessarily image-heavy, sometimes also including video and audio files. As such, users with slower internet access, or users accessing these websites via a mobile device, may have difficulty loading these content-rich pages and enjoying their full functionality. This is an issue as the ethos of the UCLA Library is that of equity and access. Scholars from communities with less web infrastructure or cheaper internet connections, for example, will not have the same access to digital collections. Further, even users with adequate internet access may find digital collections slow to load, if the page is particularly content-rich.

Software providers are continually modifying the design to improve user experience with speed and capacity. In its most recent version update, CONTENTdm, for example, added a functionality for users to choose how many results they receive from a search query on a single

---

[4] "Moving from PastPerfect to CollectiveAccess," https://collectiveaccess.org/support/index.php?p=/discussion/387/moving-from-pastperfect-to-collectiveaccess, (February 18, 2019)
[5] Price, Sara, "Collection Management Systems – Oral History in the Digital Age," http://ohda.matrix.msu.edu/2012/06/collection-management-systems/, (February 17, 2019)
[6] Butler, Nick, "Sharing Digital Collections: A Guide for Galleries, Libraries, Archives and Museums," https://www.boost.co.nz/blog/2018/10/digital-collections-galleries-libraries-archives-and-museums, October 12, 2018

page, "depending on your local connection speed and personal preference."[7] However, small design adjustments like these may serve more as temporary relief than systematic fixes. Especially as the internet becomes increasingly burdened with capacity issues with the rise of internet usage and smart phones—internet traffic increased eightfold between 2006 and 2011 alone—image-intensive applications like digital collections management systems will have to reckon with functionality and capacity.[8]

3. Copyright liabilities

As stewards of artistic materials and intellectual property, cultural heritage institutions are well versed in copyright law. Digital collections, however, increase the responsibility, exposure, and liability of copyright protection by making materials widely available to the public. Managers of digital collections thus have to be vigilant in their stewardship of materials while using digital collections management software, as the scope of their responsibility is exponentially risen when given an online platform.

Fortunately, copyright law has shaped how digital collections management systems are developed and designed. Digital collections with original or representations of works of art need to honor rights of attribution and integrity.[9] To enable this, digital collections management systems have fields that allow institutions to properly attribute works of art. Furthermore, to help prevent unlawful dissemination of materials, digital collections management systems can make materials visible, but restrict the ability for users to download the materials. Of course, there is always the possibility for a user to screenshot the image. Accordingly, digital collections management systems can also list rights information along with the digital object, making it clear to users what is protected by copyright law. With this multi-pronged approach, institutions utilizing digital collections management software are making best efforts to prevent infringement, with the onus ultimately placed on the user to follow copyright law. To enact these design protections, however, institutions must be responsible and accurate in their rights settings.

---

[7] "CONTENTdm Release Notes, March 2019," OCLC Support, February 28, 2019, https://help.oclc.org/Metadata_Services/CONTENTdm/Release_notes/2019_release_notes/100_CONTENTdm_release_notes_March_2019

[8] Blanchette, Jean-François, 2015, "Computing's Infrastructural Moment." In *Regulating the Cloud: Policy for Computing Infrastructure*, edited by Jean-François Blanchette and Christopher Yoo, 1–19. MIT Press.

[9] "COPYRIGHT OWNER'S RIGHTS," *Copyrightalliance* (blog), https://copyrightalliance.org/education/copyright-law-explained/copyright-owners-rights/, (February 24, 2019)

4. Web standards are developed by select decision makers

Web standards figure prominently in the design of digital collections management systems, as these software solutions are used to make collections accessible online. Additionally, many of these software solutions' database functions are designed to structurally rely on the Internet. CONTENTdm, for example, "uses a text-based search engine built using Internet standards and protocols" instead of being built as a relational database to facilitate faster performance for larger collections in an online setting.[10]

Accordingly, the development of web standards will inevitably impact digital collections management systems. Web standards are agreed-upon technical principles that help ensure that websites display content in the same way, no matter what browser a user might be using.[11] The World Wide Web Consortium (W3C) is a key player in this realm. W3C is an international community, with its 452 members including computing companies, corporations, financial institutions, media conglomerates, academic and research institutions, and government entities.[12] While the member list is quite extensive, members are predominantly from North America, Europe, and Asia, and largely consist of corporate, academic, and governmental stakeholders—not necessarily the public or humanitarian agencies.[13]

While the academic representation in standards-making organizations is encouraging for UCLA libraries, the distribution of W3C's members does not always serve disenfranchised or marginalized groups. As strong advocates of public service and equitable access, this might be a concern for UCLA libraries, in considering how to best serve all of its patrons. For example, in 2017, W3C published digital rights management (DRM) standards, which standardized how web video platforms allow browsers to display videos.[14] DRM technology is heavily protected by law. In the United States, for example, people who bypass DRM for legal reasons—such as making material accessible to those with disabilities—can still be prosecuted.[15] As such, some non-profits like the Internet Archive, UNESCO, and the Electronic Frontier Foundation

---

[10] "Resources," https://www.oclc.org/en/contentdm/resources.html, (February 10, 2019)
[11] "Web Standards," https://www.washington.edu/accessit/webdesign/student/unit1/module3/lesson1.htm, (February 10, 2019)
[12] "Current Members - W3C," https://www.w3.org/Consortium/Member/List, (February 10, 2019)
[13] Dickens, "Web Standards: The What, The Why, And The How," *Smashing Magazine*, https://www.smashingmagazine.com/2019/01/web-standards-guide/, January 14, 2019
[14] "Encrypted Media Extensions," https://www.w3.org/TR/encrypted-media/, (February 10, 2019)
[15] Doctorow, "Amid Unprecedented Controversy, W3C Greenlights DRM for the Web," Electronic Frontier Foundation, https://www.eff.org/deeplinks/2017/07/amid-unprecedented-controversy-w3c-greenlights-drm-web, July 6, 2017

petitioned W3C to include in the standards that members would only prosecute those who bypassed DRM to explicitly infringe on copyright. While some of W3C's members agreed with this sentiment, including the German National Library and the U.K. Royal National Institute for Blind People, ultimately such protections were not voted through, likely due to the majority corporate representation of W3C members.[16] DRM standards directly affect institutions with videos in their collections, and will be the first of many decisions from W3C that could impact digital collections—a concern if the fundamental ethos of W3C's decision-making members is at odds with the service-oriented goals of UCLA Library.

Future Trends

While collections management software has been around for several decades, its digital counterpart is a bit younger, with most starting in the mid-2000s.[17, 18] As such, its future is full of potential challenges and opportunities.

On the design front, the technology will likely continue to improve across the bar. Industry leaders like PastPerfect and CONTENTdm regularly release software updates with improvements spanning technical capacity and user experience.[19, 20] Given the high number of competitors in the market—and the need to compete on functionality—these improvements will likely continue.

That said, there are potential economic impediments to the growth and continual improvement of the technology. Many cultural heritage institutions are on unsure ground as to funding. In times of economic recession especially, when ticket sales and donations are down, institutions can be forced into cutting back on programming and staffing.[21] Even in the best of economic times, cultural heritage institutions are not necessarily flush with income as other companies or organizations might be. Accordingly, the growth of digital collections management technology might be stymied by its clients' lack of resources.

---

[16] Doctorow, "Amid Unprecedented Controversy, W3C Greenlights DRM for the Web," July 6, 2017
[17] "Omeka – Project," https://omeka.org/about/project/, (March 15, 2019)
[18] "PastPerfect Online User's Guide"http://museumsoftware.com/ppohelp/#t=Welcome.htm, (March 15, 2019)
[19] "Latest PastPerfect Museum Software Update Release Notes," https://www.museumsoftware.com/releasenotes.html, (March 3, 2019)
[20] "CONTENTdm Release Notes, October 2018," 2018, OCLC Support, https://help.oclc.org/Metadata_Services/CONTENTdm/Release_notes/2018_Release_Notes/090CONTENTdm_release_notes_October_2018
[21] Grant, Daniel, "How Do Museums Pay for Themselves These Days?" *Huffington Post*. September 7, 2012, https://www.huffingtonpost.com/daniel-grant/museum-cuts_b_1816309.html

References

Blanchette, Jean-François. 2015. “Computing’s Infrastructural Moment.” In *Regulating the Cloud: Policy for Computing Infrastructure*, edited by Jean-François Blanchette and Christopher Yoo, 1–19. MIT Press.

Butler, Nick. 2018. “Sharing Digital Collections: A Guide for Galleries, Libraries, Archives and Museums | Bigger Impact.” October 12, 2018. https://www.boost.co.nz/blog/2018/10/digital-collections-galleries-libraries-archives-and-museums.

“CONTENTdm Features.” n.d. Accessed February 2, 2019. https://www.oclc.org/en/contentdm/features.html.

“CONTENTdm Release Notes, March 2019.” 2019. OCLC Support. February 28, 2019. https://help.oclc.org/Metadata_Services/CONTENTdm/Release_notes/2019_release_notes/100_CONTENTdm_release_notes_March_2019.

“CONTENTdm Release Notes, October 2018.” 2018. OCLC Support. October 15, 2018. https://help.oclc.org/Metadata_Services/CONTENTdm/Release_notes/2018_Release_Notes/090CONTENTdm_release_notes_October_2018.

“COPYRIGHT OWNER’S RIGHTS.” n.d. *Copyrightalliance* (blog). Accessed February 24, 2019. https://copyrightalliance.org/education/copyright-law-explained/copyright-owners-rights/.

“Current Members - W3C.” n.d. Accessed February 10, 2019. https://www.w3.org/Consortium/Member/List.

Dickens, Amy. 2019. “Web Standards: The What, The Why, And The How.” *Smashing Magazine*. January 14, 2019. https://www.smashingmagazine.com/2019/01/web-standards-guide/.

Doctorow, Cory. 2017. “Amid Unprecedented Controversy, W3C Greenlights DRM for the Web.” Electronic Frontier Foundation. July 6, 2017. https://www.eff.org/deeplinks/2017/07/amid-unprecedented-controversy-w3c-greenlights-drm-web.

“Encrypted Media Extensions.” n.d. Accessed February 10, 2019. https://www.w3.org/TR/encrypted-media/.

Grant, Daniel. 2012. “How Do Museums Pay for Themselves These Days?” *Huffington Post*. September 7, 2012. https://www.huffingtonpost.com/daniel-grant/museum-cuts_b_1816309.html.

“Latest PastPerfect Museum Software Update Release Notes.” n.d. Accessed March 4, 2019. https://www.museumsoftware.com/releasenotes.html.

“Moving from PastPerfect to CollectiveAccess.” n.d. CollectiveAccess Support Forum. Accessed February 18, 2019. https://collectiveaccess.org/support/index.php?p=/discussion/387/moving-from-pastperfect-to-collectiveaccess.

“Omeka - Project.” n.d. Accessed March 15, 2019. https://omeka.org/about/project/.

“PastPerfect Online: Expand Your Audience by Providing Online Access.” n.d. Accessed February 18, 2019. http://www.museumsoftware.com/pponline.html.

“PastPerfect Online User’s Guide.” n.d. Accessed March 15, 2019. http://museumsoftware.com/ppohelp/#t=Welcome.htm.

Price, Sara. n.d. “Collection Management Systems – Oral History in the Digital Age.” Accessed February 17, 2019. http://ohda.matrix.msu.edu/2012/06/collection-management-systems/.

“Resources.” 2014. OCLC. February 11, 2014. https://www.oclc.org/en/contentdm/resources.html.

“Web Standards.” n.d. Accessed February 10, 2019. https://www.washington.edu/accessit/webdesign/student/unit1/module3/lesson1.htm.

Experience

During my time at UCLA, I valued professional experiences to the same degree as my academic studies, seeking opportunities in diverse LIS environments so I could gain as much practical experience as possible before entering the field. In addition to this experience, I also sought mentorship and professional enrichment opportunities to better orient myself within the field.

This section provides a picture of the work I have done to develop professionally, and where I see myself after the UCLA MLIS program. As such, this page includes my resume, a professional development statement that outlines my career goals, and my advising history.

Resume

Savannah Lake

savannahlake@gmail.com
858-204-6671

Metadata & Information Architecture

EDUCATION

Master of Library and Information Studies, **University of California, Los Angeles**, expected June 2020

- *Relevant Coursework:* Metadata; Digital Humanities; Computer Programming; Digital Asset Management; UX Research; Descriptive and Subject Cataloging; Content Management Systems; Data Management and Practice
- *Professional Organizations:* Special Libraries Association; README Digital Rights Group; ASIS&T (Association for Information Science and Technology)

Bachelor of Arts, University of California, Berkeley, English Literature and Italian Studies double major, May 2012

- *Honors:* Phi Beta Kappa, Departmental Citation for Top Undergraduate Student in Italian Studies, 3.9 GPA

INFORMATION MANAGEMENT EXPERIENCE

GETTY RESEARCH INSTITUTE — Los Angeles, CA

Getty Vocabularies Intern — March 2020 – Present

- Translates and integrates Italian terminology into the AAT vocabulary, developing a deep understanding indexing and taxonomy construction, maintenance, and governance while making AAT more accessible.

Metadata Intern — Oct. – Dec. 2019

- Assessed, cleaned, and reconciled legacy metadata in OpenRefine, preparing metadata for transformation to linked open data. Implemented controlled vocabularies.
- Collaborated with team members to determine data content rules that would best support user research and collection access. Created a report that documented the rules and decision-making behind them.
- Defined and documented metadata workflows in order to aid future cleaning and reconciliation efforts.
- Created a reference sheet of GREL syntax for manipulating metadata to facilitate ongoing data cleaning.

UCLA LIBRARY — Los Angeles, CA

Assistant, Digital Library Program — June 2019 – Present

- Assists in the migration and description of thousands of digital assets: creates metadata, crops and batch renames assets, uses Python scripts to generate metadata, and reviews MODS metadata files for accuracy.
- Tests digital collections websites for functionality and usability, advising on design, search, and metadata.
- Supports outreach efforts by identifying classes and professors that could benefit from collections.
- Conducts research and testing on technologies like OCR and ALTO, making recommendations for their implementation within the collections.

Public Services Assistant, Library Special Collections — June 2019 – Present

- Provides research assistance for students, faculty, and the public. Helps readers find materials, navigate the collections, and place requests through the catalog.
- Supports outreach efforts, developing instruction curriculum and coordinating paging and day-of logistics.
- Checks in readers into the reading room, communicating policies to ensure security of the collections.
- Drafted manual of reference desk procedures to better train and onboard new reference staff.
- Created documentation of technical programs to help readers access born-digital and digitized materials.
- Researched online reference technologies and workflows, advising on integration into the library's practices.

GO FOR BROKE NATIONAL EDUCATION CENTER — Los Angeles, CA

Community Archives Fellow, UCLA Community Archives Lab — Sept. 2019 – Present

- Processes archival collections, arranging and housing materials and creating finding aids on ArchivesSpace.
- Scans fragile archival material to ensure long-term access and preservation, applying metadata and appropriate naming conventions.
- Supports digitization projects with partner institutions, creating metadata for digitized items that follow project standards while also capturing complex parent/child/grandchild relationships.

Savannah Lake

UNIVERSITY OF SOUTHERN CALIFORNIA Los Angeles, CA
Research Data Intern, Schaeffer Center for Health Policy & Economics June – Sept. 2019

- Supported data reuse by creating a Drupal reference library of the center's publications and datasets.
- Designed a Drupal survey to expedite the data request process. Integrated survey design best practices to ensure accurate and consistent data collection as well as ease of use for survey takers.

THE FOWLER MUSEUM Los Angeles, CA
K-12 Student Educator Sept. 2018 – June 2019

- Led interactive, hands-on gallery tours, art workshops, and story hours for K-12 students, fostering active learning and visual literacy. Developed curricula tailored to different age groups.
- Taught groups of up to 25 students, employing group management to ensure safety and cohesion.
- Developed a comprehensive plan for the Fowler Library: identified issues in collections management and circulation practices, and provided concrete recommendations for best practices and next steps.

CORONADO HISTORICAL ASSOCIATION San Diego, CA
Collections Intern May – July 2018

- Conducted internal/external research related to readers' requests, pulling relevant archival material. Supervised in-person appointments with readers, answering questions and ensuring safe handling of archival material.
- Developed and led a training session to a team of 10 new volunteers, covering topics like archival research strategies, collections management systems, and professional visitor interaction.

ADDITIONAL PROFESSIONAL EXPERIENCE

WATER ENVIRONMENT FEDERATION Washington, DC
Books Production Specialist 2017 – 2018

- Maintained a database of book products, ensuring consistency and accuracy. Edited the organization's e-commerce website, adding SEO-friendly descriptions and product metadata to bolster web presence.
- Ensured deadlines were met, managing the progress of different departments, freelancers, and vendors. Shepherded 6 titles through production in just 3 months, a record for the organization in over 10 years.

THE WYLIE AGENCY New York, NY
Literary Assistant (2016 – 2017) 2015 – 2017

- Pitched book ideas to major publishers. Analyzed deal terms, supporting negotiations to ensure best terms.
- Read and extensively reported on manuscripts, providing editorial feedback and evaluations.
- Directly assisted clients, answering queries on the phone, in person, and by e-mail.

Contracts Assistant (2015 – 2016)

- Negotiated and drafted contracts for book and magazine deals, interfacing with major publishers.
- Analyzed dozens of contracts a day to respond to permission requests, negotiating terms and issuing licenses.

THE DEPARTMENT OF JUSTICE Washington, DC
Paralegal Specialist, Antitrust Division 2012 – 2014

- Conducted legal research and reviewed thousands of documents to further discovery. Managed the document review effort for the Anheuser-Busch/Modelo investigation, supervising a team of 20 paralegals and increasing review volume by over 50,000 documents in two months' time.

SKILLS

- **Languages:** Italian (proficient)
- **Systems and software:** Adobe Bridge and Photoshop; Confluence, and Jira; OpenRefine; Tableau; SQL; HTML and CSS; CONTENTdm, PastPerfect, and ArchivesSpace; Archivematica; Aeon; Python and Jupyter Notebook
- **Standards and thesauri:** MARC 21, RDF, AACR2, LCSH, AAT, TGN, ULAN, Dublin Core, MODS

Professional Development Statement

Metadata and information architecture are foundational to making information resources work. Whether it be records at a government agency, collections in an archive, digital assets at a corporation, or books in a library, if information resources are not described well, they will not be found. I was drawn to metadata and information architecture because as the backbone of information resources, they play such a fundamental role in how information is found and therefore used. Both metadata and information architecture require organization and logic, but also empathy, as user experience design is critical to building out usable models. I aim to pursue a career thinking through these issues of information access and management.

Experience
Previous to my MLIS, I spent six years in the workforce in the legal and publishing fields. This experience was invaluable for developing my project management and communication skills, as I worked in high-pressure environments that often required collaboration across teams. More specifically to information studies, I gained experience with records management, copyright, permissions, and search engine optimization.

During my MLIS, I sought professional opportunities that would expose me to a variety of institutions and information types, to get a better understanding of the scope of the field and my place within it. Within these opportunities, I was able to integrate what I had learned from various classes. For example, the “Metadata” and “Computer Systems and Programming” courses greatly informed my work at the UCLA Digital Library Program, where I support the description and ingest of thousands of digital assets. This work involves creating metadata models as well as using Python to generate files and automated metadata.

At the Getty Research Institute, I worked on data modeling and information architecture at a deeper level, reconciling legacy metadata to a linked open data model in one internship, and contributing to the Getty Vocabularies in another. The “Digital Humanities” and “Subject Cataloging” courses were helpful here, as my work involved significant data cleaning with OpenRefine and thinking through the implications of any elisions made in the name of standardization.

I continued this type of description work, but from the archival side, at the Go For Broke National Education Center, a community archive where I processed collections. Concepts from “Values and Communities” and “Archives, Records, and Memory” have been critical to completing this work, and understanding issues community archives are facing.

At the Schaeffer Center for Health Policy & Economics, I developed my user experience skills by designing a survey for collecting data requests. The survey transformed a complex process involving several forms into a streamlined survey that was structured to ensure accurate and consistent data collection. In addition to “Human-Computer Interaction,” critical courses for accomplishing this were “Data Management and Practice” and “Data Curation and Policy,” both of which were taught by my supervisor at the Schaeffer Center, Jillian Wallis.

I also gained experience with the front-end of information services through my work at the Fowler Museum and UCLA Library Special Collections. At the Fowler, I taught K-12 students visiting the museum, creating tours that facilitated active learning and participation. “Artifacts and Cultures” was an important course for orienting myself within the museum, and analyzing how its curation intentionally told stories. This informed the curricula I developed for my tours.

At Library Special Collections, I work on the reference desk, helping users navigate the collections and place requests through the catalog. “History of Books and Literacy Technologies” was important for deepening my understanding of special collections, while “Historical Research Methods” challenged me to think critically about primary source documents and their biases. Both of these courses have informed my reference work at Library Special Collections.

Professional Organizations and Enrichment

I am an active member of the Special Libraries Association (SLA) student chapter and Southern California chapter, and plan to remain so going into my career. In my two years with SLA, I have attended a number of different events, including tours, day-long trainings, and happy hours. At all of these activities, members have been open and helpful, sharing their experiences in the field. Because I see myself working in a digital library, or in metadata in a non-LIS environment, I feel that SLA will be especially helpful as it is meant to connect information professionals who can feel siloed in such specialized settings.

I am also a member of the Association for Information Science and Technology (ASIS&T) student chapter, through which I have attended technical trainings. ASIS&T has been a good resource for connecting with other informatics students within the program, to talk through issues and share resources. Going forward, I would like to be involved with a similar group after graduation, to ensure I keep up-to-date on LIS technologies and issues. While there is no ASIS&T Los Angeles chapter, there are alternatives like the San Gabriel Valley UX (SGVUX) group, of which I am already a member.

In addition to participating in professional organizations, I have also attended conferences, trainings, and webinars to supplement my in-class studies. This includes the Henry Stewart Digital Asset Management conference, where I met information professionals working in diverse industries like hospitality, entertainment, and retail. This year I will attend the Information Architecture Conference, SLA 2020, and Keystone Digital Humanities. For enrichment, I have attended the Library Carpentry and various data training sessions. I have also found SLA and the Association of College & Research Libraries to host helpful webinars.

Going forward, I will continue this type of engagement, as it will keep me apprised of important issues in the field and new technologies. I am especially eager to attend future conferences. Thus far, I have been fortunate enough to have secured scholarship funding for all of the conferences attended; many of these opportunities exist for early-career professionals, and I also hope that whatever institution I work for will support such enrichment opportunities.

Career Goals

While I am primarily interested in metadata and information architecture, my time doing reference and instruction work showed me that I enjoy the front-end side of information work as well. Accordingly, it is important to me to find a position that allows me to work directly with users as well. This can look different in different contexts—within a library, this could be working as a digital librarian who also leads outreach and instruction efforts; in a corporate context, this could include training people on systems and answering questions about resources.

After the program, I hope to work in a position in which I am working with metadata, including developing models, cleaning and remediating metadata, and optimizing resource discovery. I would especially enjoy doing this within an academic library as part of a digital library program, so that I would have ample opportunity to participate in reference, instruction, and outreach as well. However, I would also be happy to do this sort of work in a corporate context, perhaps in digital asset management or in content strategy. In the long term, I would like to be in a managerial, forward-thinking position, where I could set policies and mentor younger professionals.

Advising History

I am grateful for my faculty advisor in the program, Professor Jean-François Blanchette. During my first year in the program we met once a quarter to discuss my academic and professional progress. This included discussing which classes I should take, the importance of participating in professional organizations, and potential work opportunities after I finished my MLIS. During my second year in the program we met more frequently, at twice each quarter. In addition to talking about my courses and work, we also discussed my issue paper in depth. These meetings pushed me to think more comprehensively about my topic and situate it in conversation with other emerging technologies in the field.

In addition to my advisor, I regularly met with UCLA MLIS alumni Lisa Moske and Grace Lau, who I connected with through various mentorship programs. Both Lisa and Grace work in spaces of particular interest to me, with Lisa in academic technology and Grace in user experience. Through them I learned of professional development opportunities, including the Information Architecture Conference scholarship program, which I applied for and received. I am grateful for their mentorship and hope to continue the relationships after my MLIS.

And finally, I received professional advice from supervisors and colleagues at jobs. Particularly helpful was hearing about different pathways into the field and work cultures at different organizations. At the Digital Library Program, I spoke with Dawn Childress, Geno Sanchez, and T-Kay Sangwand; at the Getty Research Institute, Melissa Gill, Lily Pregill, Kelly Davis, and Matt Moore; at Go For Broke National Education Center, Gavin Do; at Library Special Collections, Neil Hodge and Matt Johnson; and at the Schaeffer Center for Health Policy & Economics, Jillian Wallis.

Issue Paper

The past ten years have seen a dramatic rise in digitization efforts in libraries. Despite this widespread interest, digitization is no small task; it requires considerable time and labor—and thus, financial resources—as skilled work is involved at nearly each stage. Unfortunately, institutions are failing to realize the full potential of these investments, chiefly due to the interface design of digital collections, which usually feature keyword search as the primary discovery model.

Keyword search is reliant on both the user's ability to know from the outset how to describe their query as well as materials' content or metadata perfectly matching said query. However, materials in digital collections do not easily fit this model. Images do not have textual content, so a keyword search for an image is wholly reliant on its descriptive metadata. And even text-based materials fail on this front, due to the limitations of the optical character recognition (OCR) technologies that enable keyword searching. OCR accuracy ratings can dip under 60% depending on the clarity of the image, the size of the font, the language of the text, and if the text is handwritten. Currently, there is no clear understanding from the user's perspective of what OCR technology is, how inconsistently it is applied across collections, and how that could affect their search results.

This paper will explore how the prominence of keyword search within digital collections combined with the limitations of OCR have failed users. This paper will include a survey of the current OCR landscape, including its capabilities and limitations. It will also identify issues that should be directly communicated to users in order to increase information literacy. And finally, this paper will explore alternatives to keyword search in digital collections, with the ultimate goal of making digital collections more navigable and useful.

The False Promise of the Keyword Search: Optical Character Recognition in Digital Collections

Savannah Lake
University of California, Los Angeles
Master of Library and Information Studies Candidate
Spring 2020

The past 10 years have seen a dramatic rise in digitization efforts in libraries. A survey conducted as far back as 2010 found that 72% of special collections at research libraries have digitization programs, with 47% participating in large-scale digitization projects (Chassanoff 459). Despite this widespread interest, digitization is no small task; it requires considerable time and labor—and thus, financial resources—as skilled work is involved at nearly each stage. This includes building and maintaining the technical infrastructure, carefully scanning (often fragile) materials to a high standard, and applying descriptive metadata to make resources discoverable.

Unfortunately, institutions are failing to realize the full potential of these investments, chiefly due to the interface design of digital collections, which usually feature keyword search as the primary discovery model. Keyword search is reliant on both the user's ability to know from the outset how to describe their query as well as materials' content or metadata perfectly matching said query. However, materials in digital collections do not easily fit this model. Images do not have textual content, so a keyword search for an image is wholly reliant on its descriptive metadata. And even text-based materials fail on this front, due to the limitations of the optical character recognition (OCR) technologies that enable keyword searching.

This paper will explore keyword search in digital collections, with an emphasis on text-heavy collections, since they especially give the false impression of effective keyword searching. While OCR technologies have been in development since the 1950s and have been commercially available for twenty years, OCR can be ineffective depending on the material it is reading (Srihari et al. 1331). Accuracy ratings can dip under 60% depending on the clarity of the image, the size of the font, the language of the text, and if the text is handwritten (Smith and Cordell 5). And without accurate OCR output, a keyword search will be unable to retrieve relevant materials. Further, sometimes an entire digital collection is OCR'd (at different accuracy rates by

item), and sometimes only certain items are OCR'd—which is all to say that there are inconsistencies in the technology itself as well as its application across a collection.

Currently, there is no clear understanding from the user's perspective of what OCR technology is, how inconsistently it is applied across collections, and how that could affect their search results. This issue is exacerbated by the prominence of keyword search in digital collections, which recalls the ubiquitous Google interface. But keyword search in a digital collection is a far cry from a Google search—Google itself has not relied on keyword search alone to retrieve results for years (Baker). This false association leads users to be overly confident in the ability of keyword search to retrieve accurate results within a digital collection.

This paper will explore how the prominence of keyword search within digital collections combined with the limitations of OCR have failed users. This paper will include a survey of the current OCR landscape, including its capabilities and limitations. It will also identify issues that should be directly communicated to users in order to increase information literacy. And finally, this paper will explore alternatives to keyword search in digital collections, with the ultimate goal of making digital collections more navigable and useful.

Current State of OCR Technology

OCR technology converts numerals, letters, and symbols into a machine-readable format by using an algorithm comprised of two elements: a feature extractor and a classifier. The feature extractor derives the features that a character possesses, while the classifier determines a character's identity by comparing it against templates of other characters (Srihari et al. 1327-28).

While this process can work well for born-digital materials or typewritten materials that are cleanly formatted, the algorithm is less effective with materials outside this mold—which

includes many of the historical materials featured in digital collections. Historical newspapers, for example, have especially low OCR accuracy on account of their complex layouts and original fonts (Chiron et al. 2). Studies have shown that errors in nineteenth-century newspapers can exceed 40%, with nearly half of the text not correctly read by OCR (Smith and Cordell 5). These problems increase with text in graphical elements; for example, OCR often struggles to recognize texts within maps (Smith and Cordell 12). Poor digitization, too, can lead to inferior OCR results, including materials digitized with earlier digital imaging equipment, materials digitized to outdated standards, and materials with substandard source media like microfilm, which is a common source for digitized newspapers (Smith and Cordell 12).

Perhaps more troubling is OCR's Western bias. While the Roman alphabet is well studied by OCR companies, other scripts like Kannada, used in India, receive little attention (Srihari et al. 1329). This bias is especially felt with Indigenous languages, which are not used as frequently to train OCR algorithms as they often have smaller datasets (Mager et al. 11). Even in languages like French and English, with corpora of 12 million OCR'd characters, 50% of errors were terms that were not in dictionaries, such as proper nouns or slang (Chiron et al. 3). These biases for Western languages and standardized words present a real challenge for institutions seeking to provide equitable access to materials, as it creates research environments that better facilitate exploration of materials from Western cultures.

Current efforts to redress these issues, while ongoing, are not tenable for most libraries, and in many ways are out of their control. Comprehensive reform would require investment from stakeholders in natural language processing, machine learning, software companies, standards committees, and libraries—essentially, consensus and commitment across industries and

institutions, which would take time and may never fully happen (Smith and Cordell 6-8). Digital collections cannot wait for this perfect world, and need to address gaps in OCR now.

Other approaches to improving OCR output include altering images, such as increasing the contrast to improve OCR legibility. Such measures only go so far, though, depending on the source material. Some scholars have built statistical models to improve OCR (Wang and Liu 16). Such work, however, requires staff with significant statistical expertise, and still does not guarantee complete accuracy. More likely for most institutions is to outsource OCR corrections. Doing this at scale, however, is extremely resource-intensive. The Australian Newspaper Digitisation Program, for example, attempted this through a two-pronged effort. First, they paid editing services to manually correct titles, subtitles, and the first four lines of each article for over 21 million newspaper pages. Then, they crowdsourced corrections to over 100 million lines of text (Smith and Cordell 10). Despite these efforts, the majority of their text remains uncorrected.

Current State of Discovery in Digital Collections

These problems with OCR technologies are compounded by the chief mode of discovery in digital collections—keyword search (Stack). Keyword search dominates user interfaces of digital collections (see Appendix A for examples), which is problematic because the search bar recalls one of the most ubiquitous information discovery platforms, Google. As such the user impulse to use a search bar is understandable, as Google is, for many, familiar and comfortable; indeed, a survey found that 97.4% of university students use Google every day (Fear 33). While a search bar in a digital collection may look like Google, it functions very differently. Google's algorithm is complex, with results dependent on not just on-page content, but also off-page factors, such as the number and quality of external links pointing to a website, paid ads, and

search history (Baker). These all work to retrieve highly personalized and developed results. That is a far cry from simple keyword matching, which is what digital collections use. Users accustomed to Google's level of accuracy may not question a keyword search, or have the framework to understand how search within different contexts work. As such, if a user does not understand OCR's limitations, they could incorrectly assume a keyword search is exhaustive.

Aside from the issues with OCR, keyword search generally has issues that interfere with a user's ability to successfully retrieve relevant resources. While successful at answering targeted questions with straightforward answers, keyword search struggles to support information-seeking with complex or speculative questions (Bates). Especially within the context of a digital collection, in which items are limited and catalogued in a specific way, it can be difficult to answer multiplex questions that may involve numerous keywords and interrelated topics without knowing the backend of how the material was catalogued (Stack).

Additionally, keyword searches discourage browsing, which can be a generative information-seeking technique. Browsing can be useful to users who are not subject-matter experts, as keyword search is necessarily a "command experience," wherein users are compelled to provide a keyword in order to begin the experience (Bates). If the results are not quite right, there is no clue or context for improving the search—as ever, the next keyword is entirely reliant on user input. This runs counter to how people naturally think, as psychology shows that recognition is easier than recall (Fedoroff and Chandler). That is, people are more likely to be able to identify their desired keyword from a selection of options as opposed to knowing the exact term from the start. Browsing fosters this more intuitive information-seeking behavior (Fedoroff and Chandler). Further, sometimes browsing is a user's explicit information-seeking

aim. For example, a survey of Dutch museum websites found that while 29% of visitors were seeking specific information, nearly just as many, 21%, visited to casually browse (Whitelaw 5).

Keyword search, then has problems very specific to OCR, misleading users into thinking they are doing exhaustive, Google searches. It also is a subpar tool for the types of complex research questions that users would likely have when using a digital collection for research.

Proposed Solution

The limitations of OCR within digital collections can be addressed through better transparency and better design, both of which will foster information literacy. User experience design is often described as being akin to infrastructure—it works best when the user does not even notice it (Halarewich). While this is certainly true in that it creates a natural, intuitive experience, such an approach does not encourage a user to think of a resource as a constructed entity. Without this awareness, a user is less likely to challenge something for its bias, or to think critically about its construction and how to navigate it. Ideally, a resource should be intuitive and navigable as well as invite the user to think about what information is and how it is produced.

Transparency

The most straightforward and cost-effective approach to addressing the limitations of OCR within digital collections is transparency. Digital collections should communicate better with their users about the composition of their collection and the mechanics of searching it. While this would in a sense "reveal the infrastructure," it is necessary information for crafting meaningful keyword searches. This means clearly identifying which collections were OCR'd, and at what level of accuracy. Providing this information enables users to be more persistent and strategic with keyword searches.

The National Archives and Records Administration published a press release for their introduction of OCR into the catalog that acknowledged OCR's shortcomings, clarifying which files were OCR'd and stating that they found "human-entered transcriptions to be more accurate than OCR" ("New Search Feature"). While this is a step in the right direction, it should be taken much further to make a true impact. This information should be within the catalog or object entries themselves, not tucked away in a press release, in order to actually reach users. Further, specific accuracy ratings should be provided, so users can make informed decisions about the collection they are reviewing. For especially text-heavy collections, in which keyword search might be primary means of entry, items and collections could even have badges with the OCR accuracy level, to readily communicate with the user (see example wireframes in Appendix B).

Metadata

Another area for addressing the limitations of OCR and keyword search is through metadata and faceted search. Faceted search allows a user to see the skeleton of how a collection was cataloged, and use filters to retrieve targeted results. Focusing attention here, over keyword search, would encourage users to explore collections in a more direct, "in the weeds" manner.

While most digital collections already have this, more effort could be concentrated here to make much-needed improvements. The California Digital Newspaper Collection (CDNC), for example, has 184 subject categories that inexplicably begin with the letter "X" ("Browse Tags"). Of these subjects, many only have one corresponding resource, which is inefficient for browsing. Further, many of the subjects include places and dates ("Browse Tags"). This is unnecessary clutter as the collection already has both location and date facets. CDNC is not alone with metadata practices that do not maximize search: UCLA Digital Collections does not reliably use controlled vocabularies, including both "Pasadena (Calif.)" and "California--Pasadena," for

example (“UCLA Library Digital Collections”); Europeana does not offer a date facet while Calisphere does not offer a subject facet (“Search”; “Search Results”); and Library of Congress does not consistently classify within facets, with prints, for example, listed in both subject and format facets (“Digital Collections”).

A potential drawback of building out metadata is that the complexity of a faceted search might discourage use. Indeed, a study on student perceptions of search tool usability found that a web-like experience is more familiar than hierarchical faceted searching (Cordes 23). Given this, it would be important to understand the users of the collection, and create straightforward yet attractive facets that would encourage use. For example, a newspaper collection could include facets relevant to news, like date and location; by contrast, an art museum’s collection could include facets art historians may find helpful, like material and technique.

Another consideration to evaluate when building out metadata is that metadata, by its nature, forces categorizing, and all of the problems intrinsic to classification. Issues of ethics within classification have garnered attention in recent years. Important scholarship includes Jonathan Furner’s work on evaluating classification schemes through critical race theory, Emily Drabinski’s engagement with queer theory and the catalog, and Marisa Elena Duarte and Miranda Belarde-Lewis’ scholarship on decolonizing classification through Indigenous knowledge organization. All informational professionals should be aware of these issues and their impact on the community, and work as inclusively as possible when cataloging.

Finally, investing in metadata is a more costly approach, requiring expert labor and in many cases remediation on work already completed. Depending on an institution’s resources, it may be necessary to prioritize certain facets that would most benefit search within the collection. Automating metadata when possible will also go far in cost-effectively describing resources.

Generous Design

An even more ambitious approach to addressing the deficiencies of OCR and keyword search is to design generous interfaces. Keyword search provides a blank slate that demands the user input a search term; by contrast, generous interfaces are rich with information, encouraging browsing and illustrating connections between materials (Whitelaw 46). Generous interfaces include depicting the collection as mosaic tiles, to facilitate browsing and communicate the scale of the collection; arranging materials by color to foster serendipitous discovery and browsing; and using maps or timelines to show where materials cluster and where gaps might be (Stack) (see Appendix C for examples). Seeing gaps in the collection could be instructive to users as to where they should put their effort in searching, replicating the experience of shelf browsing in a physical library by visually displaying the coverage of the collection (Bates; Chassanoff 463). A simple, but generous, addition to keyword search could be to offer a cluster of related terms to users once they enter in a keyword (Fedoroff and Chandler). Such approaches are more immersive than keyword search, offering diverse paths of entry.

It is important to note that generous interfaces require ample metadata and financial resources to build out. Some institutions have experimented with this—*The Queenslander* is a notable example, making their newspaper collection browsable by year, subject, and color (Whitelaw 38). However, many institutions may find this approach cost-prohibitive. Further, the technical infrastructure supporting generous interfaces can be more complex, as it can require back-end development and integration of application program interfaces (APIs). This added technical complexity requires more time and labor to both build and maintain, and may be too resource-intensive for some institutions.

Conclusion

Currently, there is a discrepancy between the prodigious amount of digitized materials and a user's ability to actually make sense of it and use it. Within this abundance of digitized material, the chief mechanism for discovery is a tiny funnel—the keyword search—that is unable to deliver rich or reliable results, in large part due to the limitations of OCR. This paper has focused on OCR and text-heavy materials, as there is less literature on this issue and search. However, there are obvious implications here for images as well, which especially suffer in keyword searches as they do not even have text that could be OCR'd—accurately or not.

Looking to the future, is artificial intelligence the answer? Currently, the algorithms are nowhere near where they need to be to deliver reliably accurate OCR across languages, fonts, and handwriting, nor are they able to consistently identify photos by subject keywords to automatically generate descriptive metadata. Materials in digital collections are often too idiosyncratic to train algorithms to this level of accuracy. Even as algorithms improve, artificial intelligence technologies will require substantial human intervention and oversight.

Libraries can, however, take steps to address these issues now, and make materials in digital collections more discoverable and thus widely used. The most cost-effective measure would be to simply inform users of these limitations, and make them active agents in their search. More costly approaches would be to build out metadata for better faceted searches, or to design generous interfaces that offer multiple, generative avenues into the collection. In all of these solutions, information professionals need to be well trained in the limitations of OCR and keyword search, so that they not only build better digital collections, but they are also better able to answer queries, supporting users in both the front end and the back end. With such measures in place, institutions can begin to realize the incredible potential of their digitization investments.

Works Cited

Baker. "A Brief History of Search Engine Optimization." *Search Engine Journal*, 26 Dec. 2017, https://www.searchenginejournal.com/seo-101/seo-history/.

Bates, Marcia J. *Neolithic Information Seeking: Designing Information Systems for Our Inner Hunter-Gatherer*. Information Architecture Summit 2018, Chicago, IL.

"Browse Tags." *California Digital Newspaper Collection*, https://cdnc.ucr.edu/?a=cl&cl=Tags.X&e=-------en--20--1--txt-txIN--------1. Accessed 23 Feb. 2020.

Chassanoff, Alexandra. "Historians and the Use of Primary Source Materials in the Digital Age." *The American Archivist*, vol. 76, no. 2, Sept. 2013, pp. 458–80, doi:10.17723/aarc.76.2.lh76217m2m376n28.

Cordes, Sean. "Student Perceptions of Search Tool Usability." *Internet Reference Services Quarterly*, vol. 19, no. 1, Jan. 2014, pp. 3–32. *Taylor and Francis+NEJM*, doi:10.1080/10875301.2014.894955.

"Digital Collections." *Library of Congress*, https://www.loc.gov/collections/. Accessed 23 Feb. 2020.

Drabinski, E. (2013). Queering the Catalog: Queer Theory and the Politics of Correction. *Library Quarterly: Information, Community, Policy, 83*(2), 94–111.

Duarte, M. E., & Belarde-Lewis, M. (2015). Imagining: Creating Spaces for Indigenous Ontologies. *Cataloging & Classification Quarterly*, (53), 677–702.

Fear, Kathleen. "User Understanding of Metadata in Digital Image Collections: Or, What Exactly Do You Mean by 'Coverage'?" *The American Archivist*, vol. 73, no. 1, 2010, pp. 26–60.

Fedoroff, Lara, and Chris Chandler. *How Search Really Works*. http://ux-radio.com/2019/04/search-really-works-guest-dr-marcia-bates/. Accessed 20 Dec. 2019.

Furner, Jonathan. "Dewey Deracialized: A Critical Race-Theoretic Perspective." *Knowledge Organization*, vol. 34, Jan. 2007, pp. 144–68.

G. Chiron, et al. *Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information*. 2017 ACM/IEEE Joint Conference on Digital Libraries, Toronto, ON. 2017, pp. 1–4, doi:10.1109/JCDL.2017.7991582.

Halarewich, Danny. "Reducing Cognitive Overload For A Better User Experience." *Smashing Magazine*, 16 Sept. 2016. https://www.smashingmagazine.com/2016/09/reducing-cognitive-overload-for-a-better-user-experience/.

Mager, Manuel, et al. *Challenges of Language Technologies for the Indigenous Languages of the Americas*. 27th International Conference on Computational Linguistics (COLING 2018), Santa Fe, New Mexico. 2018.

"New Search Feature: Optical Character Recognition (OCR)." *NARAtions*, 9 Sept. 2019, https://narations.blogs.archives.gov/2019/09/09/new-search-feature-optical-character-recognition-ocr/.

"Search." *Europeana*, https://www.europeana.eu/en/search?page=1&view=grid&query=. Accessed 23 Feb. 2020.

"Search Results." *Calisphere*, https://calisphere.org/search/?q=. Accessed 23 Feb. 2020.

Smith, David A., and Ryan Cordell. *A Research Agenda for Historical and Multilingual Optical Character Recognition*. Northeastern University, 2018.

Srihari, Sargur N., et al., editors. "Optical Character Recognition (OCR)." *Encyclopedia of Computer Science*, John Wiley and Sons Ltd, 2003, pp. 1326–1333.

Stack, John. “Exploring Museum Collections Online: Some Background Reading.” *Medium*, 6 Aug. 2018, https://lab.sciencemuseum.org.uk/exploring-museum-collections-online-some-background-reading-da5a332fa2f8.

UCLA Library Digital Collections. https://digital.library.ucla.edu/. Accessed 23 Feb. 2020.

Wang, Hsiang-An, and Pin-Ting Liu. *Towards a Higher Accuracy of Optical Character Recognition of Chinese Rare Books in Making Use of Text Model*. 3rd International Conference on Digital Access to Textual Cultural Heritage, Brussels, Belgium. 2018, pp. 15–18.

Whitelaw, Mitchell. “Generous Interfaces for Digital Cultural Collections.” *Digital Humanities Quarterly*, vol. 9, no. 1, May 2015.

Bibliography

Baker. "A Brief History of Search Engine Optimization." *Search Engine Journal*, 26 Dec. 2017, https://www.searchenginejournal.com/seo-101/seo-history/.

Bates, Marcia J. *Neolithic Information Seeking: Designing Information Systems for Our Inner Hunter-Gatherer*. Information Architecture Summit 2018, Chicago, IL.

---. "What Is Browsing—Really? A Model Drawing from Behavioural Science Research." *Information Research*, vol. 12, no. 4, Oct. 2007, http://informationr.net/ir/12-4/paper330.html.

Bell, Steven J. "Submit or Resist: Librarianship in the Age of Google." *American Libraries*, vol. 36, no. 9, 2005, pp. 68–71. JSTOR.

"Browse Tags." *California Digital Newspaper Collection*, https://cdnc.ucr.edu/?a=cl&cl=Tags.X&e=-------en--20--1--txt-txIN--------1. Accessed 23 Feb. 2020.

Chassanoff, Alexandra. "Historians and the Use of Primary Source Materials in the Digital Age." *The American Archivist*, vol. 76, no. 2, Sept. 2013, pp. 458–80, doi:10.17723/aarc.76.2.lh76217m2m376n28.

Cordes, Sean. "Student Perceptions of Search Tool Usability." *Internet Reference Services Quarterly*, vol. 19, no. 1, Jan. 2014, pp. 3–32. *Taylor and Francis+NEJM*, doi:10.1080/10875301.2014.894955.

"Digital Collections." *Library of Congress*, https://www.loc.gov/collections/. Accessed 23 Feb. 2020.

Drabinski, E. (2013). Queering the Catalog: Queer Theory and the Politics of Correction. *Library Quarterly: Information, Community, Policy, 83*(2), 94–111.

Duarte, M. E., & Belarde-Lewis, M. (2015). Imagining: Creating Spaces for Indigenous Ontologies. *Cataloging & Classification Quarterly*, (53), 677–702.

Fear, Kathleen. "User Understanding of Metadata in Digital Image Collections: Or, What Exactly Do You Mean by 'Coverage'?" *The American Archivist*, vol. 73, no. 1, 2010, pp. 26–60.

Fedoroff, Lara, and Chris Chandler. *How Search Really Works*. http://ux-radio.com/2019/04/search-really-works-guest-dr-marcia-bates/. Accessed 20 Dec. 2019.

Furner, Jonathan. "Dewey Deracialized: A Critical Race-Theoretic Perspective." *Knowledge Organization*, vol. 34, Jan. 2007, pp. 144–68.

G. Chiron, et al. *Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information*. 2017 ACM/IEEE Joint Conference on Digital Libraries, Toronto, ON. 2017, pp. 1–4, doi:10.1109/JCDL.2017.7991582.

Halarewich, Danny. "Reducing Cognitive Overload For A Better User Experience." *Smashing Magazine*, 16 Sept. 2016. https://www.smashingmagazine.com/2016/09/reducing-cognitive-overload-for-a-better-user-experience/.

Mager, Manuel, et al. *Challenges of Language Technologies for the Indigenous Languages of the Americas*. 27th International Conference on Computational Linguistics (COLING 2018), Santa Fe, New Mexico. 2018.

"More Guidance on Building High-Quality Sites." *Official Google Webmaster Central Blog*, 6 May 2011, https://webmasters.googleblog.com/2011/05/more-guidance-on-building-high-quality.html.

"New Search Feature: Optical Character Recognition (OCR)." *NARAtions*, 9 Sept. 2019, https://narations.blogs.archives.gov/2019/09/09/new-search-feature-optical-character-recognition-ocr/.

"Search." *Europeana*, https://www.europeana.eu/en/search?page=1&view=grid&query=. Accessed 23 Feb. 2020.

"Search Results." *Calisphere*, https://calisphere.org/search/?q=. Accessed 23 Feb. 2020.

Smith, David A., and Ryan Cordell. *A Research Agenda for Historical and Multilingual Optical Character Recognition*. Northeastern University, 2018.

Srihari, Sargur N., et al., editors. "Optical Character Recognition (OCR)." *Encyclopedia of Computer Science*, John Wiley and Sons Ltd, 2003, pp. 1326–1333.

Stack, John. "Exploring Museum Collections Online: Some Background Reading." *Medium*, 6 Aug. 2018, https://lab.sciencemuseum.org.uk/exploring-museum-collections-online-some-background-reading-da5a332fa2f8.

Toms, Elaine G. "Understanding and Facilitating the Browsing of Electronic Text." *International Journal of Human-Computer Studies*, vol. 52, no. 3, Mar. 2000, pp. 423–52, doi:10.1006/ijhc.1999.0345.

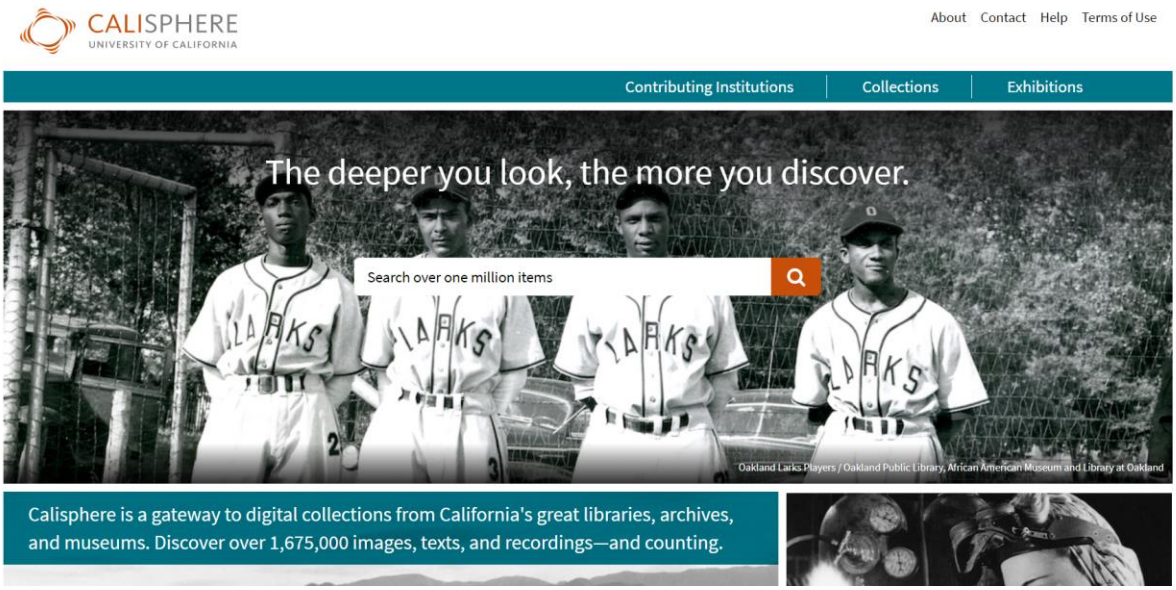
UCLA Library Digital Collections. https://digital.library.ucla.edu/. Accessed 23 Feb. 2020.

Wang, Hsiang-An, and Pin-Ting Liu. *Towards a Higher Accuracy of Optical Character Recognition of Chinese Rare Books in Making Use of Text Model*. 3rd International Conference on Digital Access to Textual Cultural Heritage, Brussels, Belgium. 2018, pp. 15–18.

Whitelaw, Mitchell. "Generous Interfaces for Digital Cultural Collections." *Digital Humanities Quarterly*, vol. 9, no. 1, May 2015.

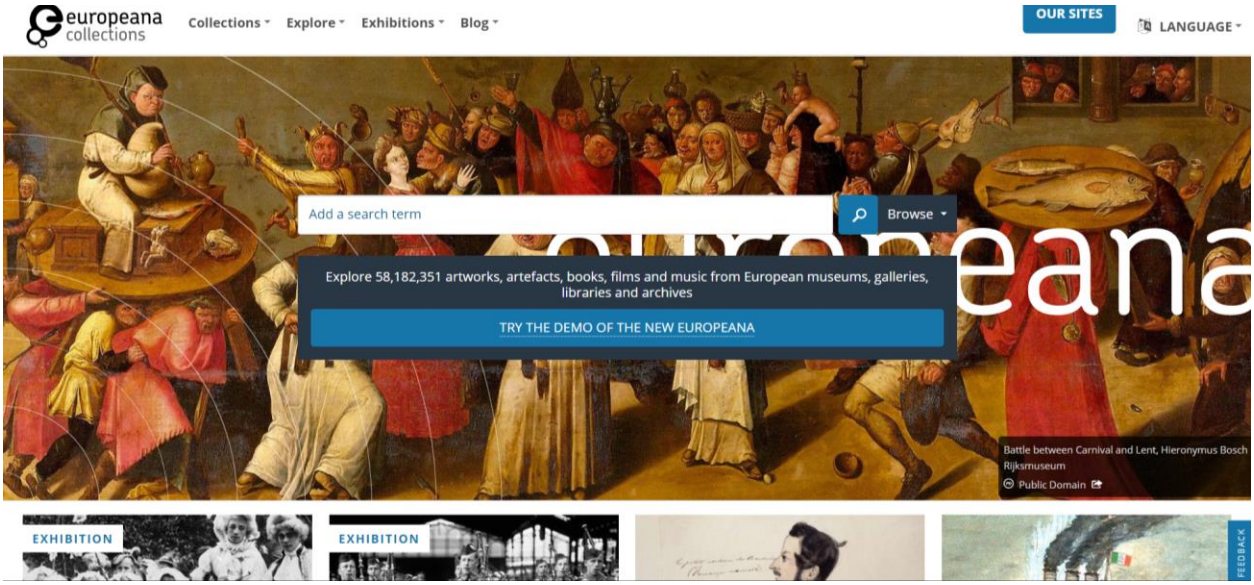
Appendix A: Digital Collections Interfaces

The homepages of digital collections often prominently feature keyword searches. Below are screen captures, all taken on February 22, 2020, that reflect the dominance of keyword search.

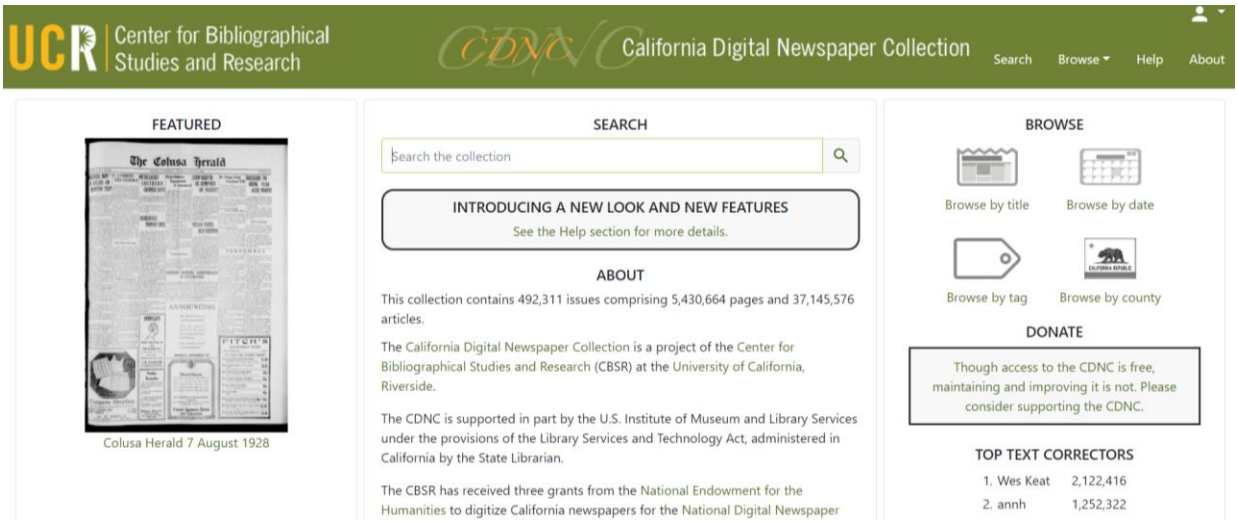


Calisphere (https://calisphere.org/)

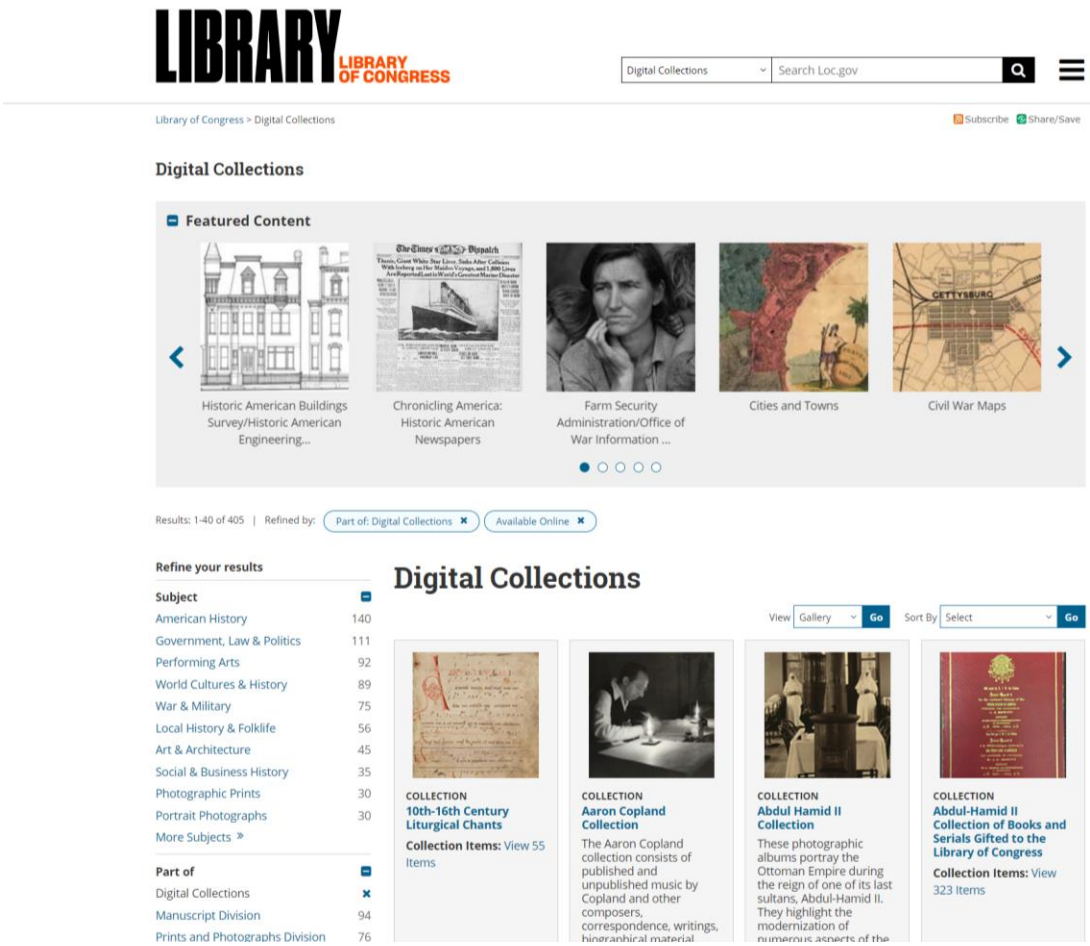
UCLA Library Digital Collections (https://digital.library.ucla.edu)



Europeana (https://www.europeana.eu/portal/en/)



California Digital Newspaper Collection (https://cdnc.ucr.edu/)

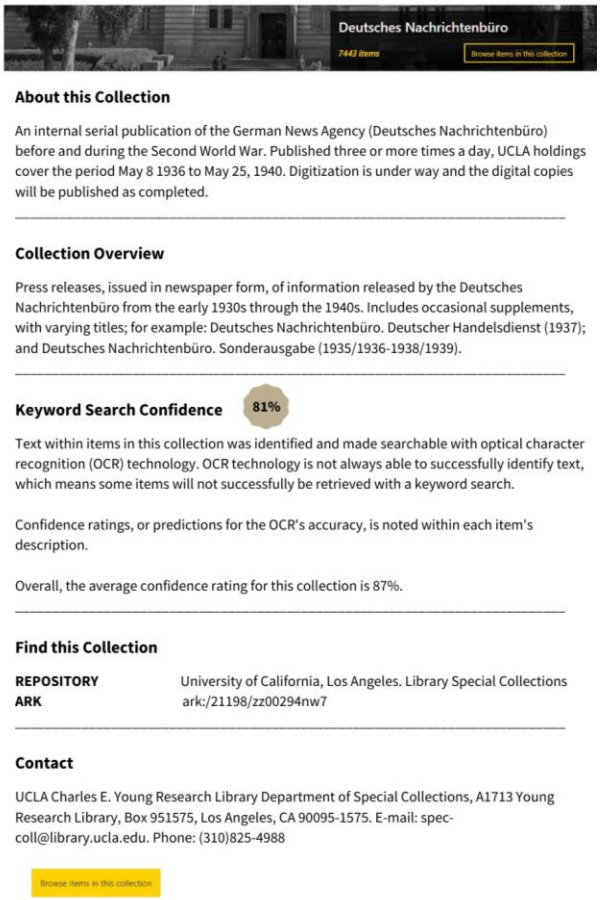


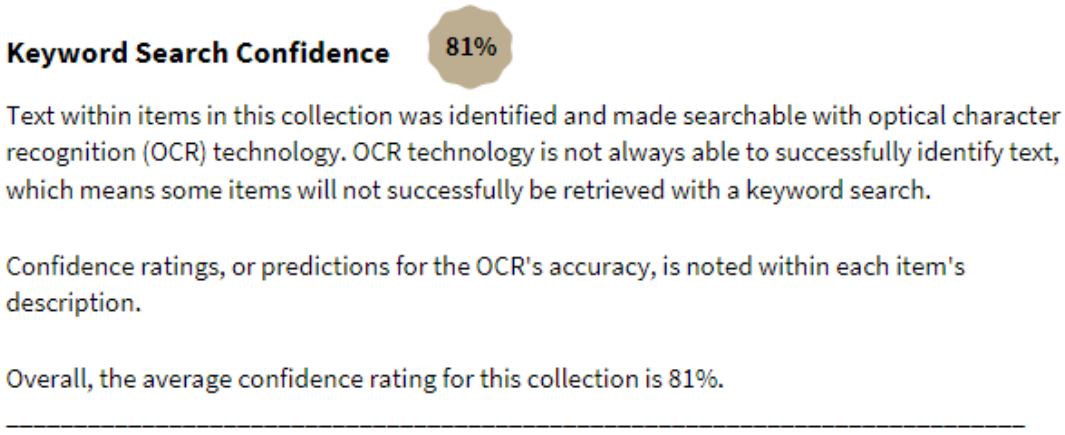
Library of Congress Digital Collections (https://www.loc.gov/collections/)

Appendix B: OCR Badge Wireframes

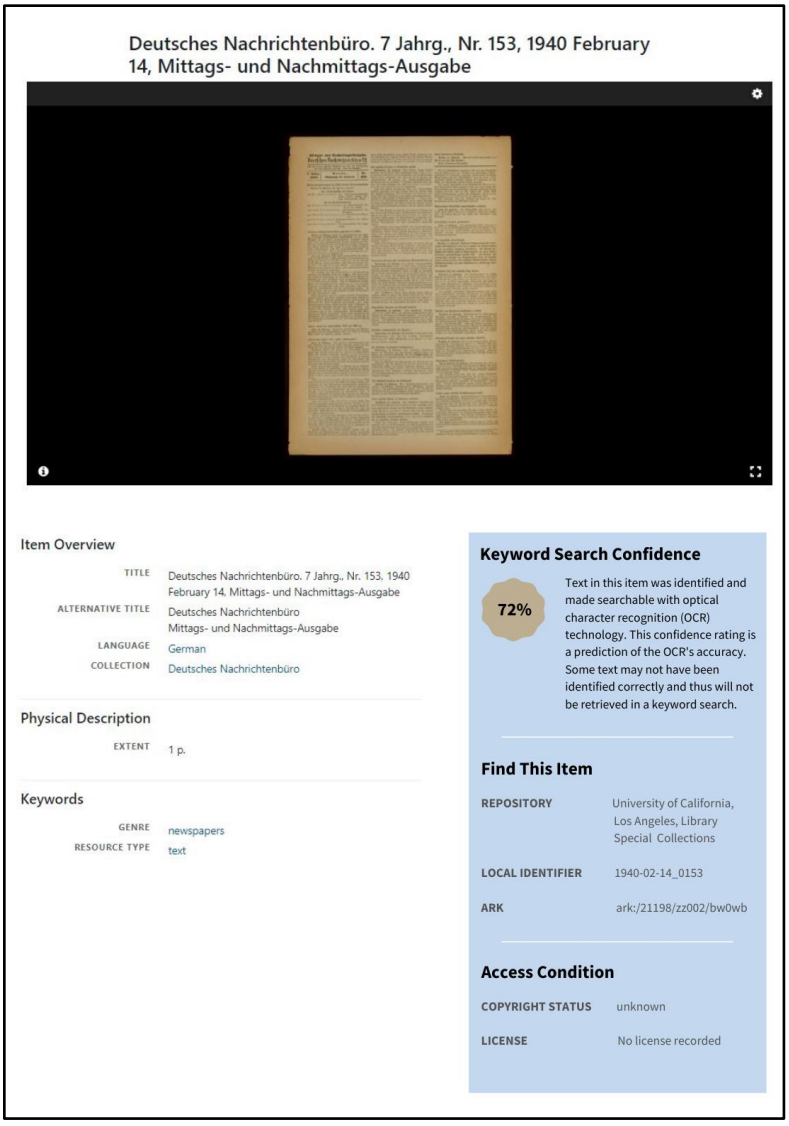
Below are wireframes of how OCR confidence ratings could be communicated to users, at the collection- and the item-level.

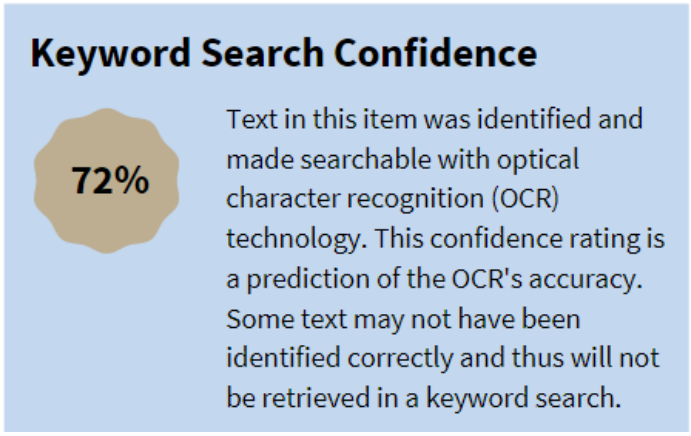
Collection-level wireframe with OCR badge





Item-level wireframe with OCR badge

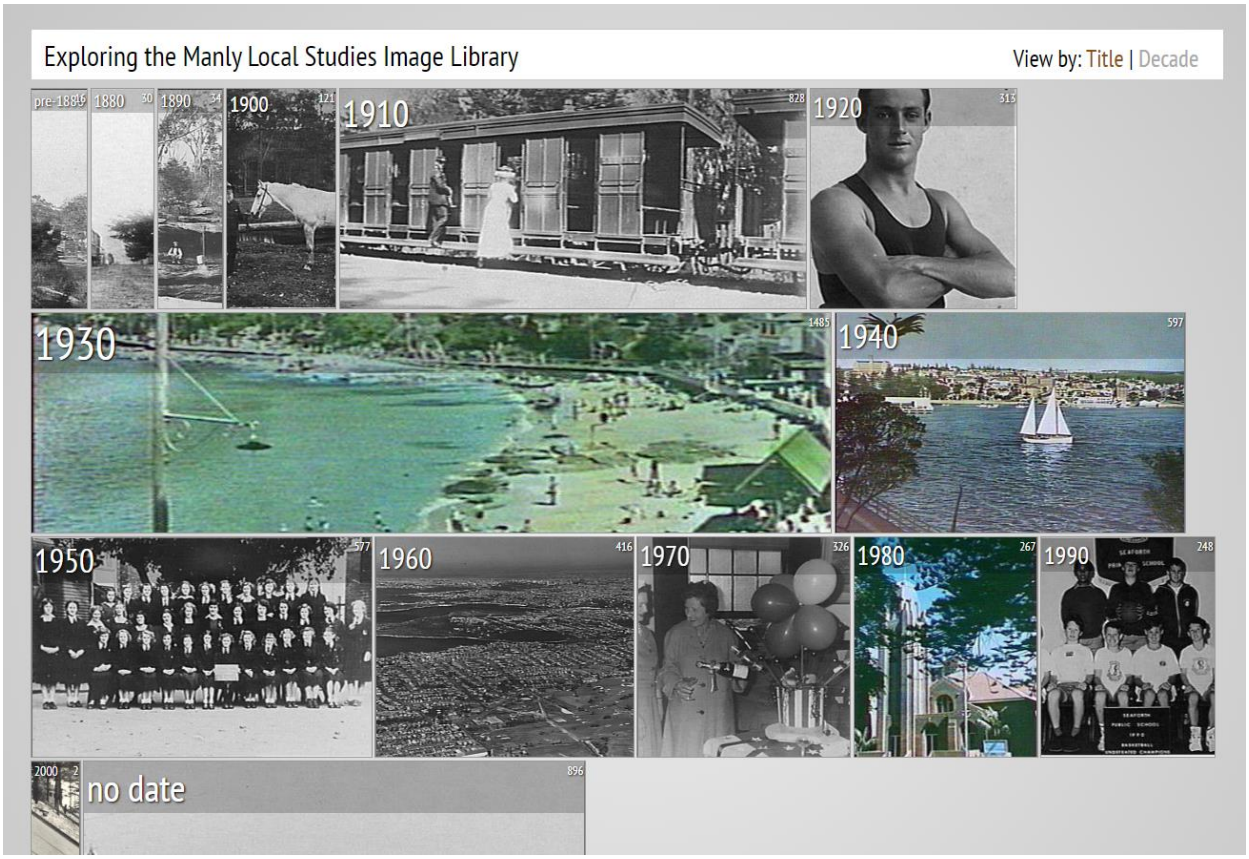




Appendix C: Generous Interface Designs

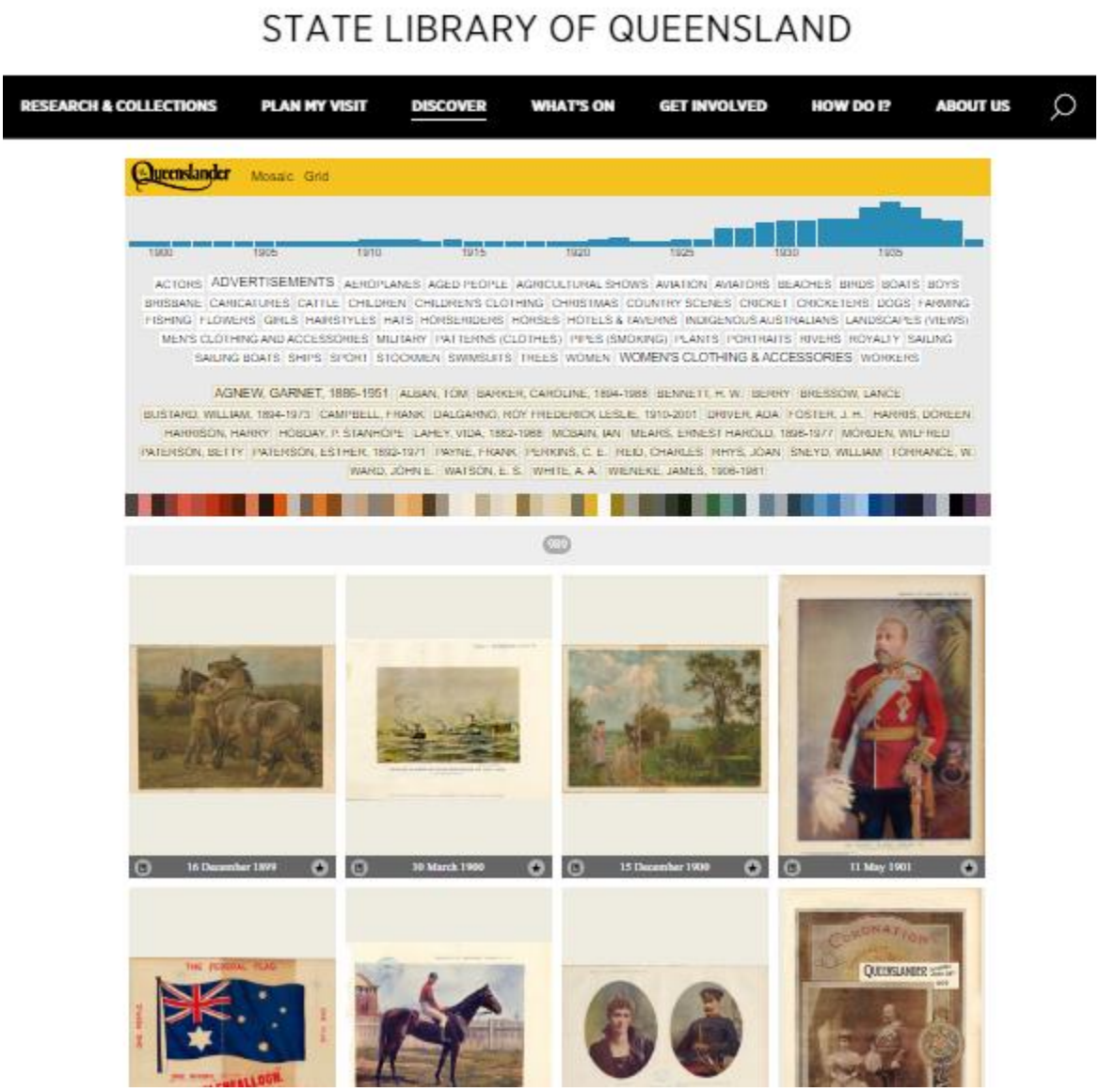
Generous interface designs work to show scale of collection and promote browsability. The following examples are described in Mitchell Whitelaw's article "Generous Interfaces for Digital Cultural Collections." The screen captures were taken on February 22, 2020.

Interface that communicates the scope of holdings by decade



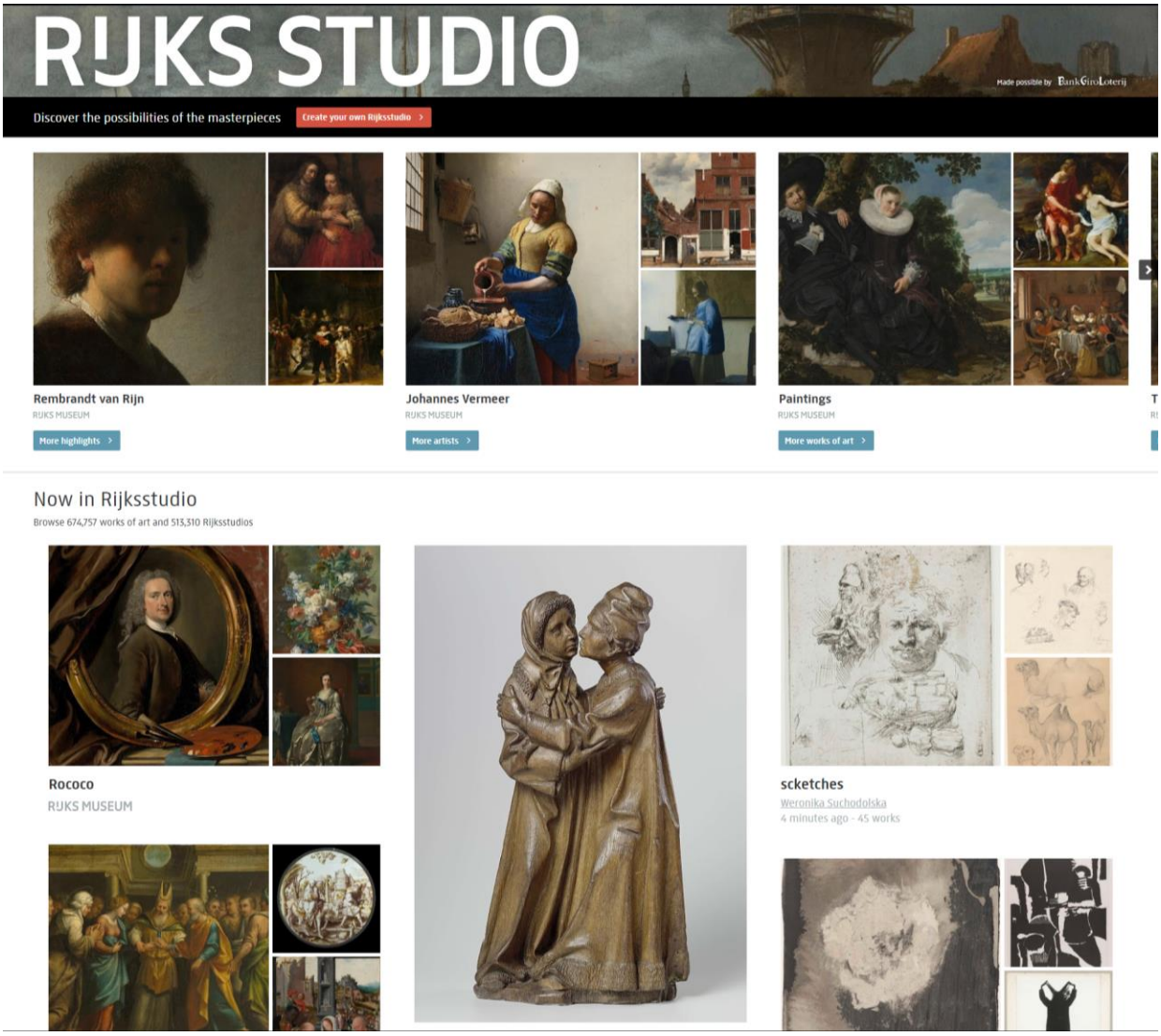
Manly Local Studies Image Library (http://mtchl.net/manlyimages/explore.html#decade)

Interface that allows browsing by subject, name, color, and date



The Queenslander (https://www.slq.qld.gov.au/discover/exhibitions/past-exhibitions/discover-queenslander#/grid)

Interface with mosaic tiles to facilitate scanning and browsing



Rijksmuseum (https://www.rijksmuseum.nl/en/rijksstudio)

Supporting Documents

My internship with the Digital Art History department at the Getty Research Institute was a valuable exercise in metadata governance and remediation. I was working on the PhotoTech project, which sought to bring a collection of over two million photographs online. The project faced some interesting challenges, largely due to the inconsistent processing and cataloging practices over the years.

My internship focused on the materials metadata, where catalogers described what materials, supports, and techniques were present in artworks. I was responsible for defining and documenting data content standards, and then remediating metadata to those standards through OpenRefine. Once the metadata was cleaned, I reconciled the metadata to the Getty Art and Architecture Thesaurus (AAT) in order to make the data more interoperable. Throughout this process, I defined and documented a workflow. I also created a reference sheet of General Refine Expression Language (GREL) syntax that I used to clean and reconcile the Photo Archive metadata. I delivered the following documentation at the end of my internship.

PhotoTech Materials Metadata

Getty Research Institute

Prepared by:
Savannah Lake
December 2019
IS 498: Internship | Professor Blanchette

Contents:

Project Summary

PhotoTech

The Photo Archive at the Getty Research Institute is an expansive collection, comprising of over two million photographic reproductions of artwork. Created in the 1970s, the archive was historically used by Getty staff and art historians to study artwork in the era before the internet. While accessing photographic reproductions of artwork is much easier now online, the items within the Photo Archive still hold strong research value, often containing provenance metadata. The PhotoTech project aims to bring the Photo Archive online, making both the images and their context more widely available for researchers.

The project faces some interesting challenges, largely due to the inconsistent processing and cataloging practices over the years. The collection as a whole has a finding aid, with varying levels of folder description. Approximately 14% of the collection has more robust item-level metadata recorded within a flat-file, command line database called the STAR database. The PhotoTech project also wishes to use computer vision to generate metadata from the stamps and handwritten notes on the photographs. The challenge will be reconciling these different sources of metadata to create an accessible and usable digital collection.

Materials Data

My internship focused on the materials metadata within the STAR database, where catalogers described what materials, supports, and techniques were present in artworks. I was responsible for defining and documenting data content standards, and then cleaning metadata to those standards through OpenRefine. Once the metadata was cleaned, I was responsible for reconciling that metadata to the Getty Art and Architecture Thesaurus (AAT), in order to make the data more interoperable. This metadata will ultimately be transformed into the linked.art linked open data model. Throughout this process, I defined and documented a workflow. I also created a cheat sheet of General Refine Expression Language (GREL) syntax that I used to clean and reconcile the Photo Archive metadata. This documentation will help with the data cleanup effort within the PhotoTech project going forward, and could even benefit metadata cleanup efforts within the Getty Research Institute more broadly.

Documentation Overview

Data Content Standards and Reconciliation Guidelines

These guidelines document the decisions made during data cleaning and AAT reconciliation. The data content standards describe how the materials data should appear, and thus how the data was cleaned and standardized. This includes rules on syntax, punctuation, and capitalization, as well as how to format anomalous terms. The reconciliation guidelines identify and explain terms that we chose not to reconcile, as well as terms that have unique reconciliation target terms.

Data Cleaning and AAT Reconciliation Workflow: Outline and Diagram

The workflow outline details the steps necessary for cleaning the materials metadata and reconciling it to AAT, providing explanations, goals, and sample GREL syntax. It also visually places the process within the broader PhotoTech Metadata Workflow diagram.

In addition to this written outline, the workflow diagram visualizes these data cleaning and reconciliation processes.

GREL Cheat Sheet

The GREL Cheat Sheet lists all of the GREL syntax used to clean and reconcile the materials data. Additionally, the cheat sheet includes use cases relevant to PhotoTech data, as well as more thorough explanations to guide any new users of OpenRefine.

Materials Reconciliation Terms and Formulas Spreadsheet

The reconciliation spreadsheet contains a list of all of the terms we are reconciling to (by material, support, and type) as well as all of the GREL formulas necessary to complete reconciliation. Any changes to these formulas can programmatically be built out there. Additionally, this spreadsheet contains all of the manual edits that will need to be executed during reconciliation.

Data Content and Reconciliation Guidelines

Contents

Overview

This document lists the data content standards for PhotoTech materials metadata, as well as decisions made with regard to reconciliation. These guidelines were drawn from historical documentation (namely the *Paintings Cataloging Worksheet Entry rules*, 1998), precedent in the data and other Getty Collections, and discussion with various members of the PhotoTech team, including Melissa Gill (metadata specialist and metadata lead for PhotoTech), Ann Harrison (special collections archivist with institutional knowledge of and experience with the STAR database), Ruth Cuadra (business applications manager and STAR database administrator), and Rob Sanderson (data architect and advisor for the linked open data model). When determining data content standards and reconciliation guidelines, the ultimate goal was to foster access and searchability while maintaining accurate description of the materials.

The data content standards apply to the first phase of the materials metadata project, in which we were cleaning legacy metadata from the STAR database. They describe how the materials data should appear in the free-text field, and thus how the data was cleaned and standardized. The data content standards include general rules on syntax, punctuation, and capitalization, as well as how to format anomalous terms.

The reconciliation guidelines apply to the second phase of the materials metadata project, in which we were reconciling terms to the Getty Art and Architecture Thesaurus. These guidelines identify and explain terms that we chose not to reconcile, as well as terms that have unique reconciliation target terms.

Below are working documents that outline some of these data decisions:
STAR data cleaning weekly meeting agendas
Reconciliation check-in meeting
Paintings database proposed changes

Data Content Standards

General rules

Capitalization
Always capitalize the first word in a field. Otherwise, do not capitalize unless the material is a proper noun.

Punctuation
Use commas only; no semicolons, periods, ampersands, or dashes.

Question marks should follow the end of a text string in parentheses.

Oil on canvas (?)
Oil on canvas?
Oil (?) on canvas

Term order
Term order should follow: medium, original support, any subsequent supports

Normalize term order when it does not change the semantics of the description. Use the most prominent occurrence if it aligns with the other content rules.

Pen and bistre or ink, watercolor (159 rows)
Pen and bistre or ink and watercolor (4 rows)
Watercolor, pen and bistre or ink (4 rows)
Pen and bistre or ink watercolor (2 rows)

For techniques, leave order as is if already incorporated semantically within the materials description.

Otherwise, if it is tacked on at the end in parenthesis, replace format with a comma.

oil on canvas (grisaille)
Oil on canvas, grisaille

Use either a comma OR an “on” for the support.

Black chalk heightened with white on gray-brown paper
Black chalk heightened with white, light brown paper
Black chalk heightened with white, on light brown paper

Multiple media values
Multiple media should be separated by commas. Do not use “and” to separate multiple media, unless it is a paired item in which the “and” connotes a relationship (eg. pen and ink, pen and wash, pen and bistre).

Pen, bistre, charcoal, and paint on paper
Pen and bistre and charcoal and paint on paper
Pen and bistre, charcoal, paint on paper

However, if there are only two media, separate them with an “and”

Red chalk and black chalk heightened with white
Red chalk, black chalk heightened with white

Repeat terms, if appropriate, in order to be exhaustive.

Black and red chalks on paper
Black chalk and red chalk on paper
Black, red chalk on paper

Specific terms

Cradles: format information about cradles at the end of the description, with a comma.

Oil on oak panel, no cradle
Oil on oak panel (w/o cradle)
Oil on oak panel (no cradle)

Lined canvas: format information about lined canvases at the end of the description, with a comma.

Oil on canvas, lined
Oil on lined canvas
Oil on canvas (lined)
Oil on canvas lined

There are many instances of “heightened with white” — leave them as is. Although not likely, there is a possibility that the work is heightened with a material other than chalk, so it is best to not to be more explicit than the data indicates. Eg. Red chalk heightened with white, gray-blue paper

Maintain the “and” for the following terms: “pen and ink,” “pen and bistre,” and “pen and wash.”

Pen, brown ink, brown wash over black chalk
Pen and brown ink and brown wash over black chalk
Pen and brown ink, brown wash over black chalk

Trois crayons: Keep trois crayons, but also add red, white, and black chalks for materials reconciliation.

Trois crayons on beige paper
Red chalk, white chalk, black chalk on beige paper, à trois crayons

Terms with additional information in the materials field: Maintain materials information only. Move other notes to the following fields.

- Move to the “object notes” field:
    - Dates for when support materials were transferred; eg. Panel, transferred to canvas (1923)
    - If the artwork is detached; eg. Detached fresco, Fresco before being detached
    - Where the artwork is located; eg. Fresco, ceiling
    - States of the artwork; eg. Ruined fresco, Damaged fresco, Original panel
    - Corrections; eg. Chalk (called pencil)
- Move to “form” field:
    - Form of the artwork; eg. Fresco fragment, Tondo
- Move to “dimension notes” field:
    - Dimensions, especially if measurements are recorded; eg. panel (octagonal)
- Move to “marks and inscriptions” field:
    - Information about inscriptions; eg. signed and dated 1735
- Keep within the “materials field”:
    - Information that identifies multiple images; eg. Dark brown and black chalk (top) black chalk (bottom), Body color (37)
    - Clarifying information about the absence of a material; eg. Tempera on linen (no gesso)

Terms to leave as-is:

- Wood: Keep “wood” as is to maintain any context, but reconcile to panel.
- Gold ground: keep variation in statements as is, and reconcile materials to “gold ground”. Eg. Tempera on gold ground panel; Tempera on panel, gold ground; Gold ground on panel
- Pierre-blanche: leave as is, without reconciling to anything in AAT. Only four records.
- Painted sketch: Not necessarily a material, but keep as is.

Reconciliation Guidelines

The emphasis for reconciliation was to reconcile with terms that would create meaningful access points in faceted searches. Accordingly, if a term would not be used often in faceted searches, or if the PhotoTech data did not have enough entries for a material to make it a meaningful faceted term, we did not reconcile the term. As all terms remain in the free-text field for materials, users will be able to locate these more obscure terms through keyword searching.

Not reconciling (too few entries, unlikely search points, or incompatible with the data model)

- Colors
- Types of wood
- Cradles
- Lined vs unlined
- Gesso
- Pierre blanche
- Anything that is not a material or a technique. This includes items classified as instruments and implements in AAT, such as pens, reed pens, and lead point.
- Oil in “oiled paper” and “heightened with oil”
- Tracing paper should not be reconciled to tracing as a technique, because sometimes people use tracing paper without tracing on it.
- “Washed” should not be reconciled to “wash” as a material, because within this dataset it describes a technique used on the support (generally, “paper”), and is an unlikely search point.
- Plywood, given the timeframe of these pieces.
- Metallic netting
- Feather

Special reconciliation cases

- When “gold” is mentioned as a material, it should be reconciled to “gold leaf.” Entries with “heightened with gold,” for example, still refer to gold leaf (not gold paint), making “gold leaf” the correct material.
- “Watercolor” and “gouache” have entries in the AAT in both the materials and technique hierarchies. They should be reconciled to their respective technique entries, as that is their more likely access points in user searches.
- To maintain both specificity for keyword searches and retrieval for faceted searches, use the following model for terms related to “panel”:
    - Wood: leave as “wood” in the materials field, reconcile to “panel”
    - Softwood: leave as “softwood” in the materials field, reconcile to “panel”
    - Board: leave as “board” in the materials field, reconcile to “panel”
- “Stone” should be reconciled to “stone (worked rock)” — Provenance Index did this as well.
- “Painted” should be reconciled to “painting (image-making)” technique as opposed to a specific type of paint material.
- For the handful of records described as "sepia" only, reconcile to "ink."
- Silk should be reconciled to “silk (textile)” instead of “silk (general, animal material),” since it is unlikely artists would be working with raw silk.
- When materials are uncertain, such as "bistre or ink,” reconcile to both for best recall.

To be determined

- Object reconciliation with pen and ink drawings and pen and wash drawings?
- “Gold ground” has been created in AAT within the materials hierarchy. If it moves to the techniques hierarchy, all instances of gold, gold leaf, and gold ground should have “gold leaf” as the material and “gold ground” as the technique.

Data Cleaning and Reconciliation Workflow

PHASE I: DATA CLEANING

While Photo Archive catalogers followed the *Paintings Cataloging Worksheet Entry rules* when cataloging in the STAR database, the materials data field is a free-text field, which means there is a substantial amount of variance in the data. The following steps outline how to standardize data in OpenRefine.

1. Load CSV into OpenRefine

When loading the data into OpenRefine, select UTF-8 as the character encoding to preserve diacritics. You will also need to update the column names, either in Excel or in OpenRefine, as data exports from the STAR database use placeholder text instead of the field name. Should you choose to update the headers in Excel, be sure to import the CSV correctly so that it maintains diacritics (Open new Excel workbook. Select Data > from text > delimited > comma > next tab > file origin: unicode UTF8 > next tab).

2. Execute GREL functions that apply to the entire dataset

Execute the following before clustering, as it will reduce the amount of discrepancies within the set:

- To sentence case (see GREL Cheat Sheet)
- Trim white space
- Any overarching spelling or syntax choices as dictated by the data content standards and AAT (eg. “gray” vs “grey”; “and” vs “&”)

3. Cluster and clean dataset

This is where the bulk of the work will come in. Clustering will group similar variants of the same term together so that you can make mass edits for the correct term. Fingerprint and n-gram fingerprint clustering will likely yield the most results, but go through all keying function options as they are able to catch different typos and aspects of clustering. Follow the data content standards when cleaning data.

4. Calculate tail to determine cleaning goals

Depending on the dataset and capacity constraints, you may not be able to clean all of the data. Speak with the project manager to determine the cleaning goal. Once the bulk of the data has been cleaned, calculate the tail of the dataset to see how many records would need to be cleaned in order to reach whatever benchmark has been agreed upon with the project manager.

You can grab counts of the data from the text facet pane, and calculate the tail in Excel or Google Sheets. Here is the spreadsheet with calculations for the PhotoTech materials data as an example.

5. Spell check

Run a spell check on all of the materials statements in Google Sheets or Excel. You will catch errors like watecolor and yelllow that clustering and ngram did not catch in OpenRefine.

6. Spot check

This step provides another level of looking at the data to help catch any remaining errors.

7. Reingest clean data into STAR

Submit the cleaned data to the STAR database administrator for reingest.

PHASE II: AAT RECONCILIATION

Now that the data is standardized and clean, you will need to reconcile it to AAT and prepare it for transformation into linked open data. To do this, you will need to get the data to a point in which individual terms can be split into different columns, as seen in this target reconciliation template, in which each material, support, and technique in a materials statement is parsed out and identified by its AAT ID.

1. Create your reconciliation CSV

Your target reconciliation spreadsheet only needs to have one instance of each material statement, as it will live outside of STAR and not with the records themselves. Accordingly, instead of working on the entire dataset with over 6,000 instances of “oil on canvas,” all you need to have is a spreadsheet of all the unique materials statements, listing “oil on canvas” just once.

To create this, request an export from the STAR database with the now cleaned materials data. View this data in OpenRefine, and from the text facet of the materials field, grab all of the unique values (see step 4 of “Phase I: Data Cleaning” for a screenshot of how to do this). Paste these into a new CSV file.

In addition to these text-string statements, create a new blank column where you will do the reconciliation work. It is important that you prepare data for reconciliation in a separate column, so that you can refer to the free-text string while implementing the following transformations and confirm that the transformations are carrying forward correctly. Further, several of the transformations require filtering down the data through a text filter on the free-text string column.

Reload this CSV into OpenRefine. As ever, whenever loading data into OpenRefine, be sure to select UTF-8 as the character encoding to preserve diacritics.

2. Prepare text strings for parsing

The ultimate goal is to parse individual values out of the text string. To do this, make the text-string as standardized as possible, leaving out any unnecessary information. It should look like a list of items, separated by a common delimiter (in our case, a comma), that you could break out into several commas. To achieve this:

Make all text lowercase

Edit cells > Common transformations > To lowercase

Remove all conjunctions and unnecessary text. This will vary by dataset, but could include with, and, on, transferred. A few examples below:

```
replace(value, " on ", ", ")
replace(value, " glued onto ", ", ")
replace(value, ", transferred to", ",")
replace(value, "transferred from ", "")
replace(value, ", and ", ", ")
replace(value, " and ", ", ")
replace(value, " heightening", "")
replace(value, " or ", ", ")
```

3. Remove irrelevant data

There is some data that, per our reconciliation guidelines, we are not reconciling. There are also some terms that need to be removed in order to successfully apply the GREL syntax for parsing data. See the "manual edits" tab of the Materials Reconciliation_terms and formulas spreadsheet for a full list of terms to remove, including when to do them (at this step, or at step 7 below).

To remove this data, you can use the GREL chomp or replace formulas, supplemented by manual edits when needed. See the GREL Cheat Sheet for example use cases.

4. Remove duplicates

Now, remove duplicates within cells. For example, a material statement of "black chalk, red chalk" needs to be reconciled to "chalk" just once.

First, add commas between each word:

```
value.replace(" ", ", ").replace(",,", ",")
```

Remove duplicates:

```
value.split(', ').uniques().join(', ')
value.replace(",,", ",")
```

While the above was useful for removing duplicates, it did erroneously insert commas between two-word terms (eg. India ink, body color). You can correct for these errors with formulas listed in the "manual edits" tab of the Materials reconciliation_terms and formulas spreadsheet. To bring up the appropriate entries, use a text filter on the "media_mat 1" column and then conduct the transformation on the "clean_media_recon" column. For example, filter "gold leaf" in the "media_mat 1" column, and then in "clean_media_recon" run Edit Cells > Transform > replace(value, "leaf", "gold leaf")

5. Assign values to material, support, and technique columns

Create columns for materials, supports, and techniques:

add columns based on current column, set to null, name: material_id, support_id, technique_id

In the material_id column, run the materials GREL expression from the "material_ID formula" tab of the Materials reconciliation_terms and formulas spreadsheet. It should look like:

```
if(cells['clean_media_recon'].value.contains("oil"),",oil paint (paint)","") +
if(cells['clean_media_recon'].value.contains("tempera"),",tempera","") ...
```

In the support_id and technique_id columns, run their corresponding formulas from this Excel.

6. Remove duplicates

Some elements repeat (for example, wood and panel both transform to panel). Eliminate these with:

```
value.split(',').uniques().join(',')
```

7. Correct incorrectly reconciled data

Manually correct terms as designated in the "manual edits" tab of the Materials reconciliation_terms and formulas spreadsheet spreadsheet. To remove this data, you can use the GREL chomp or replace

formulas, supplemented by manual edits when needed. See the GREL Cheat Sheet for example use cases.

When making these edits, be sure to review the original materials statement ("clean_media_recon" and/or "media_mat_1") to be sure you are not overcorrecting. For example, material statements with "cardboard" in them are incorrectly reconciled to both "cardboard" and "panel," because the word "board" is within "cardboard," and the board is reconciled to panel. You'll thus need to remove "panel" from reconciled supports. However, there could be a materials statement that includes both panel and cardboard—for example, "Oil on cardboard, laid on panel." These statements *should* be reconciled to both cardboard and panel.

8. Split material_id, support_id, and technique_id by separator

You will now have three columns, each with a list of the corresponding terms for that materials statement that should look something like: charcoal,chalk,pencil

In order to run the reconciliation plug in, split these columns apart so each term is in its own column:

Edit column > Split into several columns > By separator > *set to a comma*

9. Run the AAT reconciliation plugin

The full documentation of how to run the plug-in is here:
http://www.getty.edu/research/tools/vocabularies/obtain/openrefine.html

Run the plug-in on each column: material_id 1, material_id 2, etc; support_id 1, support_id 2, etc; technique_id 1, technique_id 2, etc.

Complete all steps of the documentation, including choosing the correct AAT term and creating new columns with the AAT ID. When choosing correct AAT terms as part of the reconciliation, it can be helpful to run a text facet on the original statements to confirm that all terms were correctly reconciled.

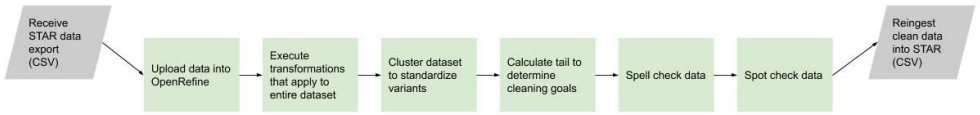
10. Create a clean AAT concordance table

Once you have reconciled all the terms, format the CSV for use as a concordance table. For reconciliation, the linked data model only requires the information as laid out in the "Transform Template" tab of the target reconciliation template.

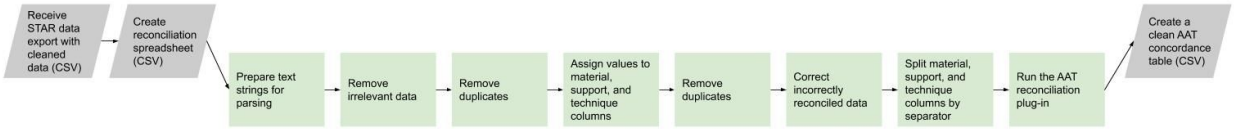
RELATED MATERIALS

This entire workflow has been diagrammed here.

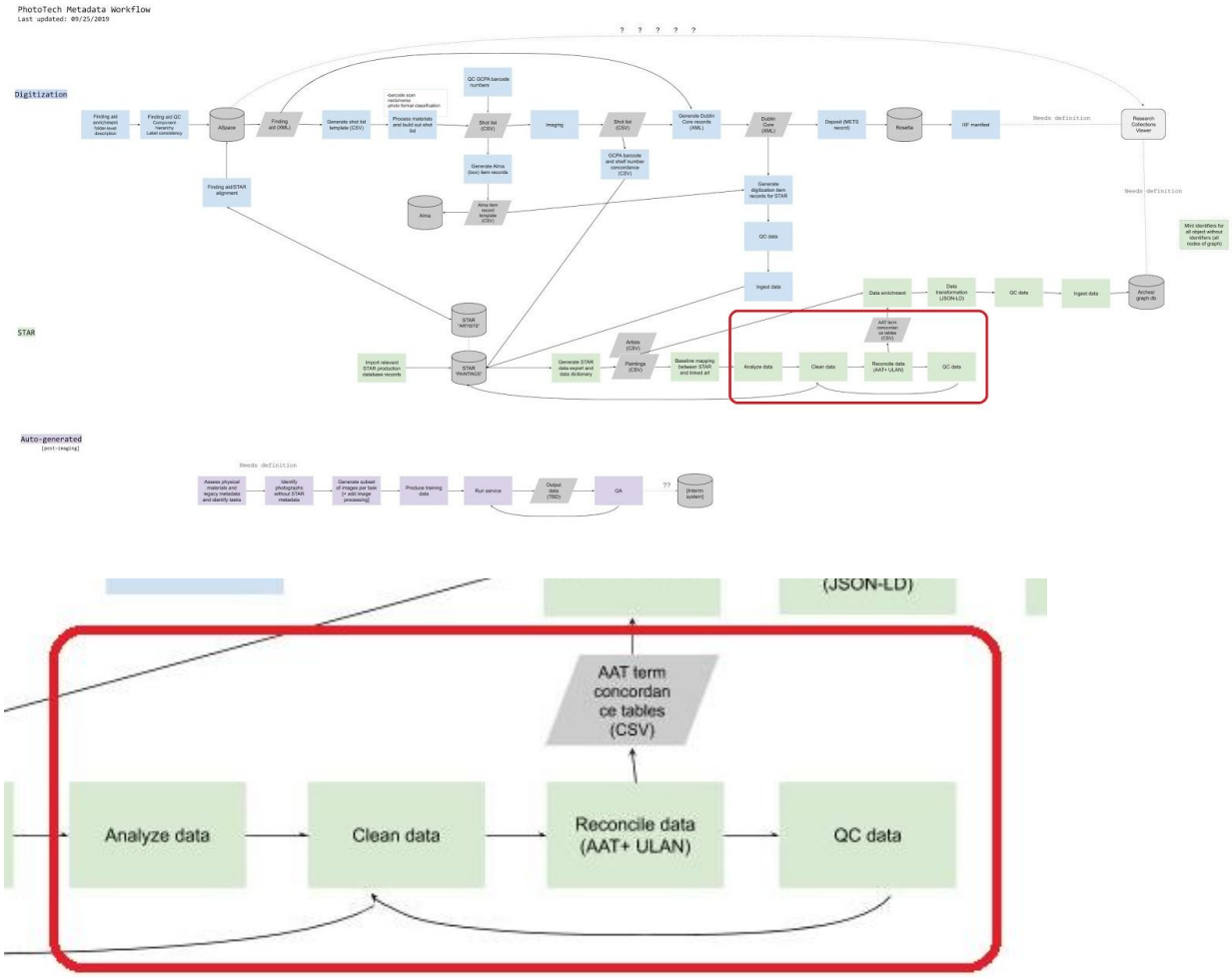
Phase I: Data Cleaning



Phase II: AAT Reconciliation

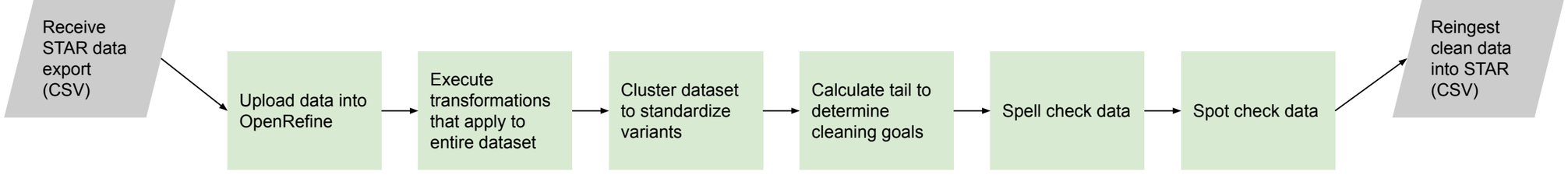


Within the broader PhotoTech Metadata Workflow diagram, these workflows fit in here:

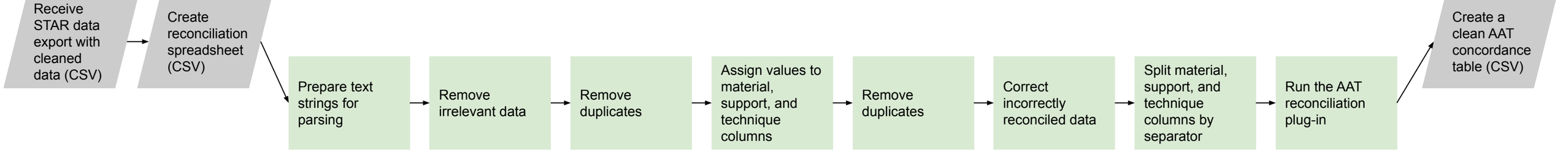


PhotoTech Data Cleaning and Reconciliation Workflow | Materials Data

Phase I: Data Cleaning



Phase II: AAT Reconciliation



GREL Cheat Sheet

Phase I: Data Cleaning

To sentence case

This function transforms text within a column to sentence case, the preferred format for PhotoTech data. This is helpful to do before clustering, to reduce clusters resulting from discrepancies in capitalization.

```
toUppercase(substring(value,0,1))+toLowercase(substring(value,1))
```

Replace

This function replaces text within a column with desired text, replacing f with r:

```
replace(value, “string f”, “string r”) OR value.replace(“string f”,“string r”)
```

This is helpful to do before clustering, to implement known standardization rules at the outset. It is also useful throughout data cleaning, when you identify trends in the data that need to be corrected.

For example, if you know that in AAT “gray” is preferred to “grey”, you can use this function to standardize.

```
Text filter > grey                  [this step is optional, but will make transformations more apparent when
                                    typing the GREL formula, so you can ensure it is implementing properly]
replace(value, "grey", "gray")
replace(value, "Grey", "Gray")      Be sure to also catch capitalized words that begin strings.
```

Further, this function can correct errors that clustering does not catch, such as ampersands.

```
Text filter > &
replace(value, "&", "and")
```

Chomp

This function removes a specified string from the end of the string if said string is there:

```
chomp(value, “string”) OR value.chomp(“string”)
```

Chomp can be an alternative to Replace when there’s a possibility of replacing the incorrect part of the string, as the Chomp function localizes any changes to the end of the string.

For example, if you would want to standardize the following strings to all end with (?):

```
Oak and panel?
Red chalk and black chalk (?)
Canvas?
```

Using Replace [value.replace(“?”, “ (?)”)] would incorrectly alter the second string, which is already correct. Instead, use chomp and add the desired text string:

```
chomp(value, “?”) + “ (?)”
```

Phase II: AAT Reconciliation

Contains and If Control

In order to use the AAT Reconciliation OpenRefine plug-in, each material within a material statement must be parsed out into separate columns. This can be accomplished through a Contains function paired with an If control.

The Contains function checks if a specified string is present, returning true or false:

```
contains(value, “string”) OR value.contains(“string”)
```

The If control evaluates whether an expression is true, implementing an expression if it is true and another expression if it is false:

```
if(expression o, expression eTrue, expression eFalse)
```

Within our use, the contains function checks a materials statement to see if it contains a certain value. If it does contain that value, it will populate a new column with that value. If it does not contain that value, it will populate a new column with nothing. For example, to check if the column named “clean_media_recon” contains the word “fresco”, and to populate the current column you are running the transformation on with “fresco” if it does and nothing if it doesn’t, use the following formula:

```
if(cells['clean_media_recon'].value.contains("fresco"),"fresco","")
```

We have programmatically constructed this formula for the materials data in the Materials reconciliation_terms and formulas spreadsheet spreadsheet.

Remove duplicates

For any materials statement, we only need to reconcile to a term once. We can remove repeated materials with the following syntax. Since each material in our formatted statements is separated by a comma, the following splits the statement by commas, extracts the unique values, and rejoins them with a comma:

```
value.split(', ').uniques().join(', ')
```

Removing irrelevant data

Part of reconciliation is removing items that could be incorrectly reconciled to an AAT term. Here, you can use the Chomp and Replace formulas again.

For example, to remove “tin” from strings like “watercolor,tin,oil” and “charcoal,tin”, you can use:

```
chomp(value, ",tin")
replace(value, ",tin,", ",")
```

Resources:

VRA Conference’s GREL Expressions for visual resources management
Common transformations (including information on replace and contains)
GREL Controls (including information on if statements)

general term	**AAT target term**	**notes**				
oil	oil paint (paint)					
bistre	bistre					
ink	ink					
charcoal	charcoal					
wash	wash					
chalk	chalk					
lead	graphite (mineral)	"black lead" captured here; when you split the terms to deduplicate, "black lead" becomes "black, lead"				
pencil	pencil (marking material)					
tempera	tempera					
graphite	graphite (mineral)					
india ink	india ink					
sinopia	sinopia					
stone	stone (worked rock)					
ground	ground					
gold	gold leaf					
gold leaf	gold leaf					
gold ground	gold ground					
pastel	pastel (material)					
targeted term	**AAT target term**	**notes**				
sepia	ink	Ignore if "sepia" precedes another material (wash, ink). If it is by itself, reconcile to ink.				

Materials reconciliation_terms and formulas

material_ID formula

general	term		target term			
if(cells['clean_media_recon'].value.contains("	oil	'),',	oil paint (paint)	',"")		if(cells['clean_media_recon'].value.contains("oil"),",oil paint (paint)","")
if(cells['clean_media_recon'].value.contains("	bistre	'),',	bistre	',"")		if(cells['clean_media_recon'].value.contains("bistre"),",bistre","")
if(cells['clean_media_recon'].value.contains("	ink	'),',	ink	',"")		if(cells['clean_media_recon'].value.contains("ink"),",ink","")
if(cells['clean_media_recon'].value.contains("	charcoal	'),',	charcoal	',"")		if(cells['clean_media_recon'].value.contains("charcoal"),",charcoal","")
if(cells['clean_media_recon'].value.contains("	wash	'),',	wash	',"")		if(cells['clean_media_recon'].value.contains("wash"),",wash","")
if(cells['clean_media_recon'].value.contains("	chalk	'),',	chalk	',"")		if(cells['clean_media_recon'].value.contains("chalk"),",chalk","")
if(cells['clean_media_recon'].value.contains("	lead	'),',	graphite (mineral	',"")		if(cells['clean_media_recon'].value.contains("lead"),",graphite (mineral)","")
if(cells['clean_media_recon'].value.contains("	pencil	'),',	pencil (marking n	',"")		if(cells['clean_media_recon'].value.contains("pencil"),",pencil (marking material)","")
if(cells['clean_media_recon'].value.contains("	tempera	'),',	tempera	',"")		if(cells['clean_media_recon'].value.contains("tempera"),",tempera","")
if(cells['clean_media_recon'].value.contains("	graphite	'),',	graphite (mineral	',"")		if(cells['clean_media_recon'].value.contains("graphite"),",graphite (mineral)","")
if(cells['clean_media_recon'].value.contains("	india ink	'),',	india ink	',"")		if(cells['clean_media_recon'].value.contains("india ink"),",india ink","")
if(cells['clean_media_recon'].value.contains("	sinopia	'),',	sinopia	',"")		if(cells['clean_media_recon'].value.contains("sinopia"),",sinopia","")
if(cells['clean_media_recon'].value.contains("	stone	'),',	stone (worked ro	',"")		if(cells['clean_media_recon'].value.contains("stone"),",stone (worked rock)","")
if(cells['clean_media_recon'].value.contains("	ground	'),',	ground	',"")		if(cells['clean_media_recon'].value.contains("ground"),",ground","")
if(cells['clean_media_recon'].value.contains("	gold	'),',	gold leaf	',"")		if(cells['clean_media_recon'].value.contains("gold"),",gold leaf","")
if(cells['clean_media_recon'].value.contains("	gold leaf	'),',	gold leaf	',"")		if(cells['clean_media_recon'].value.contains("gold leaf"),",gold leaf","")
if(cells['clean_media_recon'].value.contains("	gold ground	'),',	gold ground	',"")		if(cells['clean_media_recon'].value.contains("gold ground"),",gold ground","")
if(cells['clean_media_recon'].value.contains("	pastel	'),',	pastel (material)	',"")		if(cells['clean_media_recon'].value.contains("pastel"),",pastel (material)","")
					Combined:	if(cells['clean_media_recon'].value.contains("oil"),",oil paint (paint)","") + if(cells['clean_media_recon'].value.contains("bistre"),",bistre","") + if(cells['clean_media_recon'].value.contains("ink"),",ink","") + if(cells['clean_media_recon'].value.contains("charcoal"),",charcoal","") + if(cells['clean_media_recon'].value.contains("wash"),",wash","")

targeted	term		target term		
if(cells['clean_media_recon'].value.contains("	sepia	'),',	ink	',"")	if(cells['clean_media_recon'].value.contains("sepia"),",ink","")

Use only on the handful of cells that are only described by "sepia". All other sepia entries mention ink or wash, and will be reconciled to AAT through that.

general OpenRefine formula:

```
if(cells['clean_media_recon'].value.contains("oil"),",oil paint (paint)","") + if(cells['clean_media_recon'].value.contains("bistre"),",bistre","") + if(cells['clean_media_recon'].value.contains("ink"),",ink","") + if(cells['clean_media_recon'].value.contains("charcoal"),",charcoal","") + if(cells['clean_media_recon'].value.contains("wash"),",wash","") + if(cells['clean_media_recon'].value.contains("chalk"),",chalk","") + if(cells['clean_media_recon'].value.contains("lead"),",graphite (mineral)","") + if(cells['clean_media_rec
```

targeted OpenRefine formulas:

```
if(cells['clean_media_recon'].value.contains("sepia"),",ink","")
```

term	AAT target term	notes
panel	panel	
softwood	panel	
board	panel	
wood	panel	
cardboard	cardboard	
paper	paper	
canvas	canvas	
copper	copper	
metal	metal	
masonite	masonite	
vellum	vellum	
linen	linen	
slate	slate	
glass	glass	
tin	tin	
bronze	bronze	
cloth	cloth	
ivory	ivory	
leather	leather	
parchment	parchment	
alabaster	alabaster	
marble	marble	
burlap	burlap	
silk	silk (textile)	

	term		target term				
if(cells['clean_media_recon'].value.contains("	panel	'),'.	panel	'.*')			if(cells['clean_media_recon'].value.contains("panel"),",panel,"")
if(cells['clean_media_recon'].value.contains("	softwood	'),'.	panel	'.*')			if(cells['clean_media_recon'].value.contains("softwood"),",panel,"")
if(cells['clean_media_recon'].value.contains("	board	'),'.	panel	'.*')			if(cells['clean_media_recon'].value.contains("board"),",panel,"")
if(cells['clean_media_recon'].value.contains("	wood	'),'.	panel	'.*')			if(cells['clean_media_recon'].value.contains("wood"),",panel,"")
if(cells['clean_media_recon'].value.contains("	cardboard	'),'.	cardboard	'.*')			if(cells['clean_media_recon'].value.contains("cardboard"),",cardboard,"")
if(cells['clean_media_recon'].value.contains("	paper	'),'.	paper	'.*')			if(cells['clean_media_recon'].value.contains("paper"),",paper,"")
if(cells['clean_media_recon'].value.contains("	canvas	'),'.	canvas	'.*')			if(cells['clean_media_recon'].value.contains("canvas"),",canvas,"")
if(cells['clean_media_recon'].value.contains("	copper	'),'.	copper	'.*')			if(cells['clean_media_recon'].value.contains("copper"),",copper,"")
if(cells['clean_media_recon'].value.contains("	metal	'),'.	metal	'.*')			if(cells['clean_media_recon'].value.contains("metal"),",metal,"")
if(cells['clean_media_recon'].value.contains("	masonite	'),'.	masonite	'.*')			if(cells['clean_media_recon'].value.contains("masonite"),",masonite,"")
if(cells['clean_media_recon'].value.contains("	vellum	'),'.	vellum	'.*')			if(cells['clean_media_recon'].value.contains("vellum"),",vellum,"")
if(cells['clean_media_recon'].value.contains("	linen	'),'.	linen	'.*')			if(cells['clean_media_recon'].value.contains("linen"),",linen,"")
if(cells['clean_media_recon'].value.contains("	slate	'),'.	slate	'.*')			if(cells['clean_media_recon'].value.contains("slate"),",slate,"")
if(cells['clean_media_recon'].value.contains("	glass	'),'.	glass	'.*')			if(cells['clean_media_recon'].value.contains("glass"),",glass,"")
if(cells['clean_media_recon'].value.contains("	tin	'),'.	tin	'.*')			if(cells['clean_media_recon'].value.contains("tin"),",tin,"")
if(cells['clean_media_recon'].value.contains("	bronze	'),'.	bronze	'.*')			if(cells['clean_media_recon'].value.contains("bronze"),",bronze,"")
if(cells['clean_media_recon'].value.contains("	cloth	'),'.	cloth	'.*')			if(cells['clean_media_recon'].value.contains("cloth"),",cloth,"")
if(cells['clean_media_recon'].value.contains("	ivory	'),'.	ivory	'.*')			if(cells['clean_media_recon'].value.contains("ivory"),",ivory,"")
if(cells['clean_media_recon'].value.contains("	leather	'),'.	leather	'.*')			if(cells['clean_media_recon'].value.contains("leather"),",leather,"")
if(cells['clean_media_recon'].value.contains("	parchment	'),'.	parchment	'.*')			if(cells['clean_media_recon'].value.contains("parchment"),",parchment,"")
if(cells['clean_media_recon'].value.contains("	alabaster	'),'.	alabaster	'.*')			if(cells['clean_media_recon'].value.contains("alabaster"),",alabaster,"")
if(cells['clean_media_recon'].value.contains("	marble	'),'.	marble	'.*')			if(cells['clean_media_recon'].value.contains("marble"),",marble,"")
if(cells['clean_media_recon'].value.contains("	burlap	'),'.	burlap	'.*')			if(cells['clean_media_recon'].value.contains("burlap"),",burlap,"")
if(cells['clean_media_recon'].value.contains("	silk	'),'.	silk (textile)	'.*')			if(cells['clean_media_recon'].value.contains("silk"),",silk (textile),"")
						Combined:	if(cells['clean_media_recon'].value.contains("panel"),",panel,"") + if(cells['clean_media_recon'].value.contains("softwood"),",panel,"") + if(cells['clean_media_recon'].value.contains("board"),",panel,"") + if(cells['clean_media_recon'].value.contains("wood"),",panel,"") + if(cells['clean_media_recon'].value.contains("cardboard"),",cardboa
OpenRefine formula:							
if(cells['clean_media_recon'].value.contains("panel"),",panel,"") + if(cells['clean_media_recon'].value.contains("softwood"),",panel,"") + if(cells['clean_media_recon'].value.contains("board"),",panel,"") + if(cells['clean_media_recon'].value.contains("wood"),",panel,"") + if(cells['clean_media_recon'].value.contains("cardboard"),",cardboard,"") + if(cells['clean_media_recon'].value.contains("paper"),",paper,"") + if(cells['clean_media_recon'].value.contains("canvas"),",canvas,"") + if(cells['clean_media_recon							

term	AAT target term	notes
fresco	fresco	
silverpoint	silverpoint	
grisaille	grisaille	
brush	brush	
aquarelle	aquarelle	
etching	etching	
traced	traced	
metalpoint	metalpoint	
engraving	engraving	
varnished	varnished	
embroidering	embroidering	
à trois crayons	à trois crayons	
intarsia	intarsia	
woodcut	woodcut	
painted	painted	
gouache	gouache	
watercolor	watercolor	
body color	gouache	

	term		target term				
if(cells['clean_media_recon'].value.contains("	fresco	'),'.	fresco	'.'")			if(cells['clean_media_recon'].value.contains("fresco"),".fresco","")
if(cells['clean_media_recon'].value.contains("	silverpoint	'),'.	silverpoint	'.'")			if(cells['clean_media_recon'].value.contains("silverpoint"),".silverpoint","")
if(cells['clean_media_recon'].value.contains("	grisaille	'),'.	grisaille	'.'")			if(cells['clean_media_recon'].value.contains("grisaille"),".grisaille","")
if(cells['clean_media_recon'].value.contains("	brush	'),'.	brush	'.'")			if(cells['clean_media_recon'].value.contains("brush"),".brush","")
if(cells['clean_media_recon'].value.contains("	aquarelle	'),'.	aquarelle	'.'")			if(cells['clean_media_recon'].value.contains("aquarelle"),".aquarelle","")
if(cells['clean_media_recon'].value.contains("	etching	'),'.	etching	'.'")			if(cells['clean_media_recon'].value.contains("etching"),".etching","")
if(cells['clean_media_recon'].value.contains("	traced	'),'.	traced	'.'")			if(cells['clean_media_recon'].value.contains("traced"),".traced","")
if(cells['clean_media_recon'].value.contains("	metalpoint	'),'.	metalpoint	'.'")			if(cells['clean_media_recon'].value.contains("metalpoint"),".metalpoint","")
if(cells['clean_media_recon'].value.contains("	engraving	'),'.	engraving	'.'")			if(cells['clean_media_recon'].value.contains("engraving"),".engraving","")
if(cells['clean_media_recon'].value.contains("	varnished	'),'.	varnished	'.'")			if(cells['clean_media_recon'].value.contains("varnished"),".varnished","")
if(cells['clean_media_recon'].value.contains("	embroidering	'),'.	embroidering	'.'")			if(cells['clean_media_recon'].value.contains("embroidering"),".embroidering","")
if(cells['clean_media_recon'].value.contains("	à trois crayons	'),'.	à trois crayons	'.'")			if(cells['clean_media_recon'].value.contains("à trois crayons"),".à trois crayons","")
if(cells['clean_media_recon'].value.contains("	intarsia	'),'.	intarsia	'.'")			if(cells['clean_media_recon'].value.contains("intarsia"),".intarsia","")
if(cells['clean_media_recon'].value.contains("	woodcut	'),'.	woodcut	'.'")			if(cells['clean_media_recon'].value.contains("woodcut"),".woodcut","")
if(cells['clean_media_recon'].value.contains("	painted	'),'.	painted	'.'")			if(cells['clean_media_recon'].value.contains("painted"),".painted","")
if(cells['clean_media_recon'].value.contains("	gouache	'),'.	gouache	'.'")			if(cells['clean_media_recon'].value.contains("gouache"),".gouache","")
if(cells['clean_media_recon'].value.contains("	watercolor	'),'.	watercolor	'.'")			if(cells['clean_media_recon'].value.contains("watercolor"),".watercolor","")
if(cells['clean_media_recon'].value.contains("	body color	'),'.	gouache	'.'")			if(cells['clean_media_recon'].value.contains("body color"),".gouache","")
						Combined:	if(cells['clean_media_recon'].value.contains("fresco"),".fresco","") + if(cells['clean_media_recon'].value.contains("silverpoint"),".silverpoint","") + if(cells['clean_media_recon'].value.contains("grisaille"),".grisaille","") + if(cells['clean_media_recon'].value.contains("brush"),".brush","") + if(cells['clean_media_recon'].value.contains("aquarelle"
OpenRefine formula:							
if(cells['clean_media_recon'].value.contains("fresco"),".fresco","") + if(cells['clean_media_recon'].value.contains("silverpoint"),".silverpoint","") + if(cells['clean_media_recon'].value.contains("grisaille"),".grisaille","") + if(cells['clean_media_recon'].value.contains("brush"),".brush","") + if(cells['clean_media_recon'].value.contains("aquarelle"),".aquarelle","") + if(cells['clean_media_recon'].value.contains("etching"),".etching","") + if(cells['clean_media_recon'].value.contains("traced"),".traced","") + if(cells['clean_m							

Before splitting clean_media_recon into material, support, and technique columns (step 3 of the reconciliation workflow)

Remove from clean_media_recon

any cradles and their materials

lead point

Additionally, you'll need to prevent overactive reconciliation—for example, reconciling "tin (metal)" to the term "tinted".
The following terms have potential for this problem, and should be removed as designated (steps 3 and 7 of the reconciliation workflow).

term	incorrect attribution	remove BEFORE splitting recon_data (step 3)	remove AFTER splitting recon_data (step 6)
wood	plywood	plywood	
tin	tinted, painting, netting, writing, tint, acquatint	tinted, netting, writing, tint, acquatint	painting
metal	metallic		
oil	oily, oiled, heightened with oil	oily, oiled, heightened with oil	
cardboard	panel (because board is reconciled to panel, and "board" is in "cardboard")		panel
ink	pink, pinkish	pink, pinkish	
gold	golden, gold leaf when it should just be gold ground	golden	gold leaf when it should just be gold ground
ground	background	background	
gold ground	ground		ground
washed	wash	washed	
metalpoint	metal		metal

Add back two-word terms after removing duplicates (step 4 of the reconciliation workflow)

term	formula	notes						
black lead		no need >> captured in "lead", see "material_terms" sh	replace(value, "	lead	", "	black lead	")	replace(value, "lead", "black lead")
body color	replace(value, "body", "body color")		replace(value, "	body	", "	body color	")	replace(value, "body", "body color")
india ink	replace(value, "ink", "india ink")		replace(value, "	ink	", "	india ink	")	replace(value, "ink", "india ink")
gold leaf	replace(value, "gold", "gold leaf")		replace(value, "	leaf	", "	gold leaf	")	replace(value, "leaf", "gold leaf")
gold ground	value.replace(" ground,", "").replace("gold", "gold ground")	DO NOT use this formula on entries of just "ground"	replace(value, "	ground	", "	gold ground	")	replace(value, "ground", "gold ground")
à trois crayons	replace(value, "crayons", "à trois crayons")		replace(value, "	crayons	", "	à trois crayons	")	replace(value, "crayons", "à trois crayons")

After splitting recon_data into material, support, and technique columns (step 7 of the reconciliation workflow)

Remove colors and other terms that have been confused for materials or supports:

ivory	"ivory paper" should not be reconciled to "ivory"
gold	"gold pen" should not be reconciled to "gold leaf"
copper	no confusion with copper in this data set, but others have the potential to conflate color with material here
bronze	no confusion in bronze this data set, but others have the potential to conflate color with material here
tracing paper	should not be reconciled as a technique