# The Semantic Impact of Metadata Translation:
# A Case Study of the *Chinese in California* Virtual Collection Across Digital Platforms

Jackson Huang

Submitted in partial fulfillment of the Masters Thesis Option
Completed April 24th, 2019

University of Michigan School of Information
Advisors: Assistant Professor Patricia Garcia, Associate Professor Paul Conway
Committee Member: Gwendolyn Calvert Baker Collegiate Professor Lisa Nakamura

**Introduction**

My senior year of college, I attended a talk[1] by filmmaker Shola Lynch about the archival research she conducted in making her documentary film *Free Angela and All Political Prisoners*. She recounted how she became the expert on archival materials about this time and subject because while there were substantial existing materials, records were scattered across archival institutions and difficult to locate. She recalled finding a video of Angela Davis at a film archive at UCLA, described only with the title "Interview with Student at Protest" and a date. Aside from being inaccurate, as Angela Davis was a professor and not a student at UCLA -- and a high profile and controversial one at that, this title, Lynch noted, was not particularly useful for someone looking for materials related to Angela Davis, communism, women's history, Black history or any of the subjects related to the clip. While that interview was safe in the archive, the item was not actually accessible because it was not labeled in a way that would allow users to locate it. While chronological organization of materials may make sense for an archivist or experienced academic researcher, it can be difficult for other users to navigate. Topical or subject guides to archival materials also present their own challenges, since archives want users to be able to interpret the meaning of materials for themselves, archivists may not have sufficient time or subject specialist knowledge to organize materials by subject, and controversial or complex topics may be difficult to frame or label appropriately. Lynch emphasized that despite Angela Davis's relative fame, the difficulty of finding relevant archival records was related to the the broader invisibility of materials on the history of marginalized communities, and especially multiply marginalized communities, within archival institutions.

This conference presentation illustrates the difficult and time-intensive process of accessing archival materials related to the history of marginalized communities. The research reported in this thesis emerges from a personal and professional desire to recover the submerged voices of marginalized communities within the archives that lay dormant in part because they have been invisiblized by archival processes, and how these problems are complexified by digital

[1] Shola Lynch, "From Filmmaker to Curator: The Politics of Archiving," The Humanities and the Arts in the Integrated Knowledge University: Opening the Archive, (Philadelphia: University of Pennsylvania, 13 September 2014).

archives. The meaning of an archival record is heavily dependent on context, and records rely on metadata -- information about the record such as title, creator, subject access points, and description[2], rather than simply the content of the record such as the text or image, to establish meaning and help a user find what they are looking for. Archival metadata thus plays and integral part as both a source and potential solution to what I have termed "semantic harm" -- the harm caused by the certain perspectives from the realm of knowledge production. Inadequate description and insufficient access points for archival materials related to marginalized communities hurts both those communities who do not have access to their records and the field of information and to the field of archives and information itself, which loses the perspectives of members of those communities who choose not to enter the field because they do not see the relevance of archiving to their own experiences.

One of the major challenges with archival projects related to marginalized communities is that while they are powerful, they are often time-intensive, narrow in scope, and limited in reach. Building trust with communities and gaining subject specialty knowledge takes time, and the specificity of local context means that community archives tend to emphasize depth rather than breadth[3]. While there are several community archives that have made their materials available online[4], it can still be difficult for a users to find what they are looking for, since each community archive exists separately, with little interaction -- a problem that archives generally deal with, since there is no archival field wide catalog like WorldCat for libraries. In an effort to address these access issues, recent projects like Europeana and the Digital Public Library of America (DPLA) serve as aggregators for digital archival objects, allowing users to access materials from a wide range of institutions simultaneously. These aggregators are understood as

---

[2] Greg Bak, "Continuous Classification: Capturing Dynamic Relationships among Information Resources," *Archival Science*, 2012: 299.

[3] Trish Luker, "Decolonising Archives: Indigenous Challenges to Record Keeping in 'Reconciling' Settler Colonial States," *Australian Feminist Studies* 32, no. 91–92 (April 3, 2017): 108–25; Sue Mckemmish, Shannon Faulkhead, and Lynette Russell, "Distrust in the Archive: Reconciling Records," *Archival Science* (2011); Brett Lougheed, Ry Moran, and Camille Callison, "Reconciliation through Description: Using Metadata to Realize the Vision of the National Research Centre for Truth and Reconciliation," *Cataloging & Classification Quarterly* 53, no. 5–6 (July 4, 2015): 596–614.

[4] Michelle Caswell, Marika Cifor, and Mario H. Ramirez, "'To Suddenly Discover Yourself Existing': Uncovering the Impact of Community Archives," *The American Archivis*t 79, no. 1 (June 1, 2016): 56–81.; Sonia Yaco, Ann Jimerson, Laura Caldwell Anderson, and Chanda Temple, "A Web-Based Community-Building Archives Project: A Case Study of Kids in Birmingham 1963," *Archival Science* 15, no. 4 (December 1, 2015): 399–427.

simply improving the accessibility and reach of materials, boosting content without really changing it, which makes it relatively easy and safe for institutions to join. Aggregators especially have the potential to improve access to materials related to the histories of marginalized communities. For these communities, whose materials are often scattered throughout various larger institutions and smaller projects, aggregate platforms provide an opportunity to access everything in a single consolidated location -- at least hypothetically.

This thesis follows one particular archival collection, the *Chinese in California* virtual collection, across its various formats and platforms, from collection-specific to aggregate websites, in order to examine the ways that the technical transformation of the records and their corresponding metadata as a part of archival process affects their finability and intelligibility. In Lynch's experience with archival footage of Angela Davis, the records are there in the archives, the lack of appropriate metadata makes the records difficult to find and understand, silencing the voices of marginalized communities by not providing appropriate description and thus access to the records. Topical collections around the history of marginalized communities, such as *Chinese in California*, thus seek not only to consolidate relevant materials in a single place but to provide additional historical and archival context so the records can be better understood. In order for an archival collection to move from a single-institution to and aggregate platform, materials and their corresponding metadata much be transformed to meet the ingest standards of the new platform. However, a change in format and platform for both records themselves and their corresponding metadata can cause a significant semantic impact on records related to marginalized communities, which rely very heavily archival descriptions of context in order to establish meaning. In order to examine the impact of archival aggregation on the history of marginalized populations, this thesis addresses two research questions: 1) How does the translation from single institution repositories to aggregate platforms change representation of archival records and collections and thus their context and meaning?; and 2) how do digital processes like online hosting and aggregation that seek to place information in a flattened interoperable network interact with the context-specific correctives to improve visibility and understanding for collections on marginalization communities?

## Literature Review

*Provenance and the Archival Bond: What Makes a Record*

Before discussion can proceed, we must start be understanding what we mean when we say "archive" or "record." I choose to build off of Caswell's (2016) distinction between the usage of "archives" as used in the archival field to denote actually existing institutions related to records, and "the archive" in the humanities sense, which is more of a conceptual space or sociocultural structure that materializes and determines a particular society's ability to understand its own history[5]. While archives certainly fit into larger sociocultural structures and play an important role in collective memory, archives themselves can be understood specifically through Caswell's definition as "collections of records, material and immaterial, analog and digital (which, from an archival studies perspective, is just another form of the material), the institutions that steward them, the places where they are physically located, and the processes that designated them 'archival.'"[6] In Caswell's understanding there are four meanings of an archive or archival: 1) a collection of records (with specific understandings of collection and record), 2) an institution that stewards collections of records, 3) the physical repository for those collections, and 4) the processes that make materials archival, which are different from the processes that define other types of materials like bibliographic or archaeological materials. This fourth meaning of archival, and the complexities of this meaning, will be the focus of this thesis.

Integral to understanding what differentiates archives from libraries, museums, and other institutions that hold knowledge is what exactly makes something archival -- in other words, what makes something into a record. While all materials collected in an information system exist in relation to humans, records are created by individuals or organizations as representations or by-products of their activities[7], and consequently have the potential to serve as evidence of that

---

[5] Michel Foucault, *The Archaeology of Knowledge*, translated by A. M. Sheridan Smith (London and New York: Routledge, 2002): 129; Jacques Derrida, *Archive Fever: A Freudian Impression*, (University of Chicago Press, 1996).

[6] Caswell, Michelle, "'The Archive' Is Not an Archives: On Acknowledging the Intellectual Contributions of Archival Studies," *Reconstruction* 16, no. 1 (2016): 3.

[7] Sir Hilary Jenkinson, *A Manual of Archive Administration* (London: Percy Lund, Humphries & Co., 1937); Geoffrey Yeo, "Concepts of Record (1): Evidence, Information, and Persistent Representation," *The American Archivist* 70, no. 2 (2007): 334.

activity[8]. In other words, records contain information about and must be understood within the context of their creation and use. Records are used in two primary ways, first by the creator or members of the organization to facilitate the function of the organization, during either the active or archival parts of the record's life cycle, or by researchers trying to understand how the record fit or fits into the function of an organization, once the record has moved from active use into an archival repository. Bak expands on this understanding of an archival item as "a document in the context of its relationships [, ...] context that, read together with the content (for example, a document) constitutes the record".[9] When a record moves from active use into an archives, this new contextualization also becomes a part of the record itself -- the record includes both the document itself and any accompanying metadata added over time, such as the creator, date of creation, information surrounding the transfer of the record to the archive, and so on. Indeed, each additional use of the record adds more contextual information, changing the meaning slightly with each use, termed by Ketelaar as its "semantic genealogy"[10].

In fact, the context of the record is traditionally considered more important -- to the archivist at least -- than the content of the record because of the principle of provenance[11]. The Society of American Archivists defines provenance as the "origin or source of something" or "information regarding the origins, custody, and ownership of an item or collection".[12] In its simplest form, provenance refers to the direct source of acquisition of a particular collection -- which person or organization transferred, donated, or sold a particular set of records to the archives, and where relevant, the chain of custody -- who owned the materials before the donor, which helps determine the authenticity and rightful legal custodianship of the materials. Provenance is important because if records help establish evidence of activity, it is vital to understand whose activity it establishes evidence of. Along with establishing where or who records were received from, provenance also helps determine other aspects of organization and

---

[8] Furner, Jonathan. "Conceptual Analysis: A Method for Understanding Information as Evidence, and Evidence as Information." Archival Science 4 (2004): 233-265.
[9] Bak, "Continuous Classification: Capturing Dynamic Relationships among Information Resources," 299.
[10] Eric Ketelaar, "Tacit Narratives: The Meaning of Archives," *Archival Science* 1, no. 2 (2001): 138.
[11] Jenkinson, *A Manual of Archive Administration*.
[12] Society of American Archivists, "Provenance," *A Glossary of Archival and Records Terminology*, https://www2.archivists.org/glossary/terms/p/provenance.

access. For example, the terms of the transfer may include legally-mandated retention and disposal schedules or restrictions governing access and reproduction of materials.

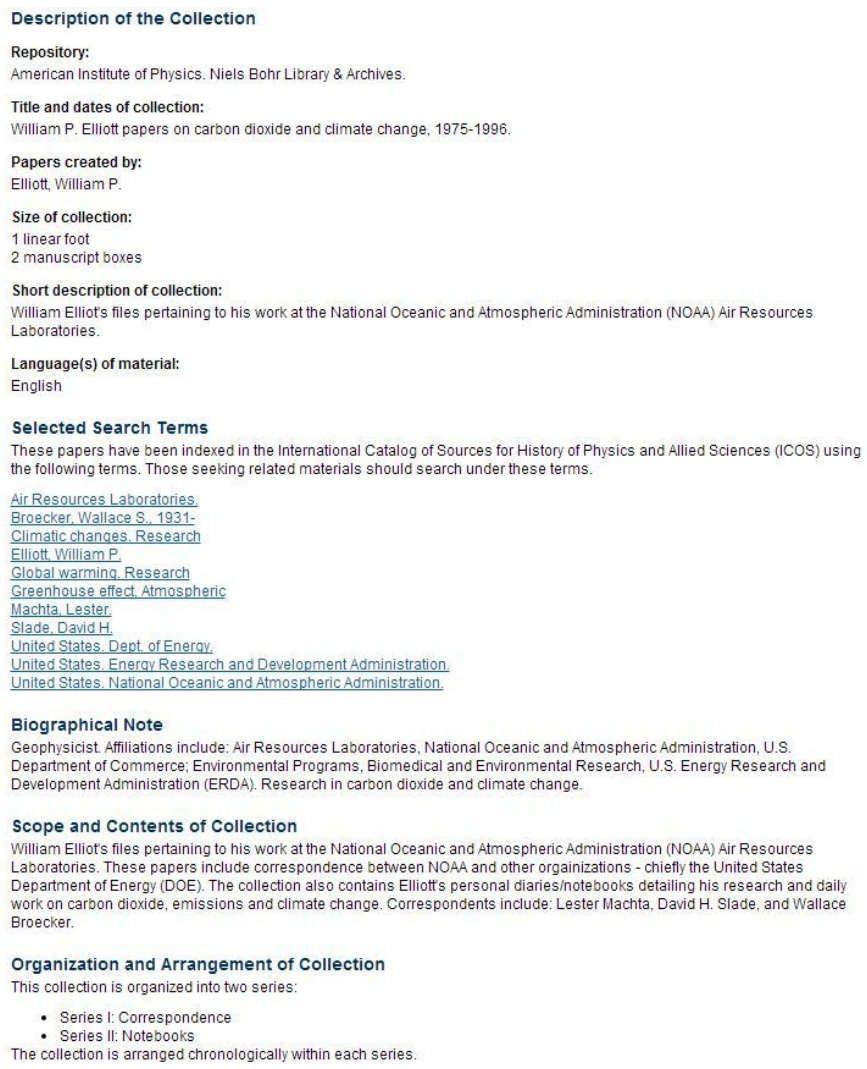
Provenance gives rise to two additional principles in archival arrangement -- *respect des fonds* and original order[13]. A *fonds* is a set of records that share a single provenance: they were received by the archive from the same source -- for example, the papers of a single individual or the administrative records of a particular company. The *fonds* should be kept together rather than broken up or interspersed with records from another *fonds* to organize materials by subject or type. A collection of records or "an archive" is usually a single *fonds*. Additionally, records should be kept in the order that the creator had arranged them in -- original order. Original order is meant to preserve the evidential nature of the records, so that the user can best understand how a particular record fit into the context of the *fonds* as a whole from the perspective of the individual or organization that produced the record. The agency or active role of the archivist is minimized, and the archivist is understood to be preserving the state of the records as they were used by the organization and received by the archives, leaving the work of interpretation and meaning-making to the user of the records.[14] Following original order, the *fonds* may additionally be be divided into series, sub-series, and files based on the existing structure within the *fonds*, such as different record types or uses, time periods, or media. The degree of granularity of arrangement depends on the extent or size of the collection, the importance of the collection to the archival institution, and staff time or expertise.

However, archival arrangement and description, like any other forms of knowledge organization, are undertaken by humans for specific purposes. Archival processes around arrangement and description are designed to meet the particularities of records and to facilitate the goals of archivists and record managers, as well as others who may have a use for or relation to records, such as donors, subjects, researchers, and members of the public. Aside from the supporting access and use of records, archival "description is tasked with placing the record in its archival bond, documenting the record's reliability, supporting presumptions of authenticity, demonstrating the transparency of archival processes, and supporting the archive's obligations to

---

[13] Samuel Muller, Johan Adriaan Feith, and R. Fruin, *Manual for the Arrangement and Description of Archives: Drawn up by the Netherlands Association of Archivists*, translated by Arthur H. Leavitt (New York: H. W. Wilson, 1968).
[14] Jenkinson, *A Manual of Archive Administration*: 146.

its creators."[15] Because the context of the records within the function of an organization is more difficult to understand once the records are removed for their original context of use and into the archives, description can clarify and explain how the records fit into functions that lead to their creation and to elucidate the various stages in the record's life cycle. The emphasis on understanding records within their original context of creation and use leads to high emphasis on the centrality of provenance for archival arrangement and description. For example, the Archives, Personal Papers, and Manuscripts (APPM), a content standard for archival cataloging promoted by both the Library of Congress and Society of American Archivists, "recognizes the primacy of provenance in archival description. This principle holds that that significance of archival materials is heavily dependent on the context of their creation, and that the arrangement and description of these materials should be directly related to their original purpose



**Description of the Collection**

**Repository:**
American Institute of Physics. Niels Bohr Library & Archives.

**Title and dates of collection:**
William P. Elliott papers on carbon dioxide and climate change, 1975-1996.

**Papers created by:**
Elliott, William P.

**Size of collection:**
1 linear foot
2 manuscript boxes

**Short description of collection:**
William Elliot's files pertaining to his work at the National Oceanic and Atmospheric Administration (NOAA) Air Resources Laboratories.

**Language(s) of material:**
English

**Selected Search Terms**
These papers have been indexed in the International Catalog of Sources for History of Physics and Allied Sciences (ICOS) using the following terms. Those seeking related materials should search under these terms.

Air Resources Laboratories.
Broecker, Wallace S., 1931-
Climatic changes. Research
Elliott, William P.
Global warming. Research
Greenhouse effect, Atmospheric
Machta, Lester.
Slade, David H.
United States. Dept. of Energy.
United States. Energy Research and Development Administration.
United States. National Oceanic and Atmospheric Administration.

**Biographical Note**
Geophysicist. Affiliations include: Air Resources Laboratories, National Oceanic and Atmospheric Administration, U.S. Department of Commerce; Environmental Programs, Biomedical and Environmental Research, U.S. Energy Research and Development Administration (ERDA). Research in carbon dioxide and climate change.

**Scope and Contents of Collection**
William Elliot's files pertaining to his work at the National Oceanic and Atmospheric Administration (NOAA) Air Resources Laboratories. These papers include correspondence between NOAA and other orgainizations - chiefly the United States Department of Energy (DOE). The collection also contains Elliott's personal diaries/notebooks detailing his research and daily work on carbon dioxide, emissions and climate change. Correspondents include: Lester Machta, David H. Slade, and Wallace Broecker.

**Organization and Arrangement of Collection**
This collection is organized into two series:

- Series I: Correspondence
- Series II: Notebooks

The collection is arranged chronologically within each series.

Figure 2. Screenshot of the finding aid for the William P. Elliott papers on carbon dioxide and climate change at the Niels Bohr Library & Archives, American Institute of Physics, showing *fonds*-level description. Information, such as extent or size of the collection, selected search terms, biography, and scope and content, apply to the collection as a whole, rather than any particular item within the collection.

[15] Stacy Wood, Kathy Carbone, Marika Cifor, Anne Gilliland, and Ricardo Punzalan, "Mobilizing Records: Re-Framing Archival Description to Support Human Rights," *Archival Science* 14, no. 3 (October 1, 2014): 400.

and function."[16] Archival description generally occurs at the level of the collection, providing contextual information about the records and their creation, such as who created them, how they came to the archive, when they were created, and a general outline of how the materials are organized, rather than at the level of each individual item, like for library materials. This information is listed in a finding aid, which provides a high-level overview of the background and organization of the collection, so that the user gets a sense of what kind of material is contained in the collection and what boxes and folders they can request to access particular records. Finding aids, such as the finding aid for the William P. Elliott papers on carbon dioxide and climate change[17] shown in figure 2, contain information for the *fonds* as a whole, though they may also contain more detailed information about particular series or even items. A more in-depth exploration of archival metadata as a record of the archival processes of arrangement and description within the context of knowledge organization can be found later in this literature review.

*The Sociopolitics of Provenance: Addressing Harm in the Archives*

However, while *respect des fonds* and original order exist an an ideal of archival arrangement and description in practice across much of the archival profession, there has also been significant work done to make visible the role and impact of archivists and archival processes in actively shaping the meaning of records. If a record consists of both its content and its context, then the recontextualization of the record from the function of the organization to the archive will inevitably change the record itself. Additionally, archives often acquire limited and incomplete record sets and physically and conceptually reorganize and rehouse records into new series, such as separating out oversize materials or media with special preservation needs. Even when archives attempt to minimize their impact, there is an integral "distinction—a conceptual one, at least—between original order and archival order"[18]. The concept of provenance has been

[16] Steven L. Henson, "The First Shall Be First: APPM and Its Impacts on American Archival Description," *Archivaria* 35 (Spring 1993): 67.
[17] Guide to the William P. Elliott papers on carbon dioxide and climate change. American Institute of Physics, Niels Bohr Library & Archives, College Park, MD 20740, USA. Screenshot from Archives @ PAMA, Region of Peel, "How Do Archivists Describe Collections? (or, How to Read a Finding Aid)".
[18] Brien Brothman, "Orders of Value: Probing the Theoretical Terms of Archival Practice," *Archivaria* 32 (Summer 1991): 85.

similarly challenged and expanded[19] to include not only the direct source of the records but the broader cultural and political setting that lead to the creation of the record.[20] The "origin" and "ownership" here are expanded to consider the various people and communities who have touched and shaped the record beyond the creator, including the subjects of records, the active role of the archival institutions in the acquisition certain materials and not others, the archivist's job in processing, arranging, describing, and preserving records, and users who reinterpret and revalue the materials.[21] In this way, community archives and thematic collections can be understood not as abandoning but enriching and complexifying notions of provenance.

Archives as a profession and discipline emerged around the management, preservation, and retrieval of government records, both to improve the efficiency of state agencies and also to ensure transparency and accountability in their mission of serving the public good.[22] However, when archivists manage records for institutions or states that harm marginalized communities, archival processes become an integral part of enabling that harm. As Wood et al explain,

> archival description and recordkeeping more broadly have been identified by both archival scholars and government inquiries as key agents in the oppression, marginalization, silencing, alienation and traumatization of individuals and communities that have been involved in social justice and human rights movements, for example, through how acts and victims are classified, euphemized, or submerged.[23]

While archivists and record managers may see the source and content of the records as separate from their role simply as organizers and preservationists of information, by facilitating the work of repressive and even genocidal governments and institutions, they play an active and integral role in the violence and material harm that comes to victims of those regimes.

Additionally, materials about and from marginalized communities are also often collected without permission or under false pretenses, such as documentation of sacred rituals that are not

---

[19] David Bearman and Richard Lytle, "The Power of the Principle of Provenance," *Archivaria*, no. 21 (1986): 14.
[20] Jennifer Douglas, "Origins: Evolving Ideas about the Principle of Provenance," in *Currents of Archival Thinking*, edited by Terry Eastwood and Heather MacNeil, (Santa Barbara, CA: Libraries Unlimited, 2010): 24.
[21] Tom Nesmith, "Still Fuzzy But More Accurate: Some Thoughts on the 'Ghosts' of Archival Theory," *Archivaria* 47 (1999): 146; Laura Millar,"The Death of the Fonds and the Resurrection of Provenance: Archival Context in Space and Time," *Archivaria* 53 (2002): 12-13.
[22] Luke Gilliland-Swetland, "The Provenance of a Profession: The Permanence of the Public Archives and Historical Manuscripts Traditions in American Archival History," *The American Archivist* 54, no. 2 (Spring 1991): 160-175.
[23] Wood et al., "Mobilizing Records," 398.

meant to be recorded or items and artefacts taken by force during colonial campaigns[24]. Consequently, Indigenous communities often distrust archives that hold stolen or unfairly taken materials and may feel that these institutions are actively complicit their community's ongoing disenfranchisement by the settler state. McKemmish et al explain, to many Aboriginal and Torres Strait Islander peoples in Australia, records containing information about their communities are considered Aboriginal and Torres Strait Islander records, and control and access of these materials should belong to Aboriginal and Torres Strait Islander communities. On the other hand, archives received these materials as part of the transfer of a governmental department's records and view these materials as belonging to the government and maintained by the archive in service to the governmental department.[25] In these cases, the archives' reliance on provenance through a donor agreement or record retention schedule is interpreted as a legitimizing or sanctioning of the harmful actions of the individual donor or governmental department over restitution or justice on the behalf of the community. This view of archival institutions as choosing to continue enabling harm is shared by Indigenous and otherwise politically marginalized communities in other contexts beyond Australia, as well, including the US.

Because archival description usually centers provenance, it tends to provide more information about the creators than subjects of records. However, this privileging of the *fonds* as the fundamental unit of an archives means that certain demographic groups who have traditionally been more able to collect and deposit archives have their voices centered, at the expense of others. Traditional notions of provenance rely on a narrow understanding of intellectual and property ownership that privileges those with the social and political power to amass records: government and corporate bodies, as well as wealthy, powerful individuals from privileged social groups. Because of the role archives play in shaping historical narratives by claiming to provide evidence of what did or did not happen, attempts at addressing the historical and ongoing role of archives in social and political structures of harm require reconceptualizing the framework for the arrangement and description of archival materials[26]. The role that archival

[24] Mckemmish, Faulkhead, and Russell, "Distrust in the Archive."
[25] Mckemmish, Faulkhead, and Russell, "Distrust in the Archive," 219.
[26] Wood et al, "Mobilizing Records."

description plays in clarifying and preserving records as evidence of action and injustice must be balanced against uncritically presenting records as they are received, which can undermine the gravity of human rights violations by legitimating euphemistic framing of violence or dehumanizing descriptions of victims. By making explicit information about the creator but not the subject of records through the centering of the context of records creation rather than the contents, norms of archival description center those with power and further reinforce the control and objectification of marginalized communities by the state, even if that is not the archivist's intent. Because records are usually collected and donated by those in positions of power, attachment to the archival concept of original order, where records are kept in the arrangement that they were received by the archive, further marginalizes communities who are more often the subjects of records as opposed to their creators because it frames them as they were seen by others -- and as other -- instead of how they see and describe themselves.

Aside from creating distrust in archival institutions, the lack of accurate, respectful, or findable records on marginalized communities limits understanding and acceptance of the active roles that these communities play in broader social and historical processes, both for members of these communities and for others. Caswell, Cifor, and Ramirez articulate an adaption of symbolic annihilation for archival contexts through "the absence or misrepresentation of their communities in archival collection policies, in descriptive tools, and/or in collections themselves."[27] Symbolic annihilation is the invisibilization of certain communities in the public sphere, where the lack of representation implies that those particular communities have not made meaningful contributions to the broader community or that they do not exist at all. Reductive or non-existent portrayals of certain communities in the archives means that there is no evidence of their contributions to their broader community and thus implies that they do not belong as meaningful shapers of the society they live in. This negatively impacts their self-perception, their ability to see themselves as historical agents and as researchers of their own history, and for other archivists, historians, and academics to gain a full understanding of the community. Symbolic annihilation is not simply the absence of collected records on certain subjects or

[27] Caswell, Cifor, and Ramirez, "'To Suddenly Discover Yourself Existing.'"

communities, but also includes reductive, harmful, culturally incompatible, or untrue representations of those subjectivities within the archives.

While the reliance on provenance has caused significant harm to marginalized communities, context is important, since context is what keeps the record in its archival bond -- understandable as a part of human processes and activities. Consequently, work done in the archival field to address the harm caused by historic and ongoing archival processes seeks to preserve evidence of these harms, while also working collaboratively with communities to build archives with culturally sensitive description that meets their needs. McKemmish articulates this ongoing process of archival contextualization and recontextualization as the records continuum model[28], where "archiving processes preserve [records] as evidence of [social and organizational] activity by disembedding them from their immediate context of creation, and providing them with ever broadening layers of contextual metadata."[29] In this understanding of the archival process, the archivist does not solely preserve the record or collection exactly as received -- an impossibility anyways, since depositing the materials in the archive already shifts context and meaning -- but takes an active role in making explicit the processes and changes in context that help a user understand a record within its broader social, political, and historical contexts through practices of archival processing, arrangement, and description. Social justice and reparative archival work should thus be understood as enriching rather than abandoning archival understandings of provenance and context.

Leveraging archives for social justice requires not only making materials related to marginalized communities available, but shifting how users engage with and conceptualize the materials and what those materials evidence, which means changing the archival processes around description and arrangement. This requires archivist to acknowledge their active role in determining what information is relevant to understanding records, and the impact that those decisions have on how different types of users interact with and encounter materials. For example, despite no such field existing in the original records, the Documentation Center of Cambodia decided to add an additional searchable metadata field for ethnicity to

---

[28] Frank Upward, "Modelling the Continuum as Paradigm Shift in Recordkeeping and Archiving Processes and Beyond," *Records Management Journal* (December 2000).
[29] Sue McKemmish, "Placing Records Continuum Theory and Practice," *Archival Science* 1 (2001): 335-336.

autobiographical accounts from victims of the Khmer Rouge, when such information was explicitly stated within the record. This additional bit of metadata, added by archivists, had a huge impact on the interpretation of scholars and activists of the war crimes of the Khmer Rouge regime as a genocide, and played an integral role as evidence of such in courts and tribunals[30].

Approaches that balance the knowledge of the community with the expertise of the archivist can be understood as addressing semantic harm using cognitive justice[31], where "different conceptions of knowledge can co-exist, and that Western knowledge can and should treat non-Western knowledge equally."[32] The post-custodial approach to archival ownership is one such strategy shifts the context of ownership and agency, especially over culturally sensitive materials, back to the community. Traditionally, ownership of archival materials is understood as fairly straightforward: records produced by an individual or organization belong to that individual or organization until complete physical and intellectual ownership are transferred to the archival institution through a deed of gift. However, in the post-custodial model, archivists and archival institutions are able to provide their professional expertise around preservation, arrangement, and management without acquiring physical custody or intellectual ownership of the records. This allows the creators of the records greater flexibility over the description and access of their materials, while the archival institution is allowed greater insight into the context of record creation and use through sustained relationship-building with the record creator[33]. While the archival institution does have less power to use its own discretion around records management, it builds trust with the community that their needs and expertise are valuable and valued.

Community archives and thematic collections also shift whose voices are centered in archives by adjusting whose records are kept and where. By shifting records creation and custodianship from formal governmental and academic institutions to community organizations,

---

[30] Michelle Caswell, "Using Classification to Convict the Khmer Rouge," *Journal of Documentation* 68, no. 2 (2012): 162–84.
[31] Shiv Visvanathan, *A Carnival for Science: Essays on Science, Technology and Development* (London: Oxford University Press, 1997).
[32] Heather Moulaison Sandy and Jenny Bossaller, "Providing Cognitively Just Subject Access to Indigenous Knowledge through Knowledge Organization Systems," *Cataloging & Classification Quarterly* 55, no. 3 (April 3, 2017): 131.
[33] Christian Kelleher, "Archives Without Archives: (Re)Locating and (Re)Defining the Archive Through Post-Custodial Praxis," *Journal of Critical Library and Information Studies* 1, no. 2 (July 7, 2017): 14-15.

archival processes occur within the community who is the subject of the records, which are often communities that have been historically institutionally excluded from archives and knowledge production. In contrast to the function-based description and arrangement conventional for governmental and corporate records[34], records in community archives tend to contain rich, textually-dense description of the story or history of the subject or communities depicted. Participation in community archives tends to prioritize depth rather than breadth, recognizing that norms that prioritize volume, such as more product, less process (MPLP), can have a disproportionate negative impact on materials concerning marginalized communities, since they tend to be a part of other collections and are easily subsumed. Collection, arrangement, description, and access tends to be tailored to the needs and context of the particular community the archive is working with, and is based on ongoing negotiation and relationship-building over time.

When community archives and thematic collections focus on a particular topic and collect materials from diverse sources, they are not holding content as more important than context, but rather understanding social and political factors and identities like race, gender, sexuality, and language, to be integral to the context of human and organization activity and thus of record creation and collection. Community archives expand understanding of provenance and origin to include the broader sociopolitical circumstances and context that led to the creation of a particular record. However, the large investment of time and energy, as well as the unique protocols based on context, mean that while projects like the South Asian American Digital Archive (SAADA)[35] and *Kids in Birmingham, 1963*[36] are historically rich, they tend to be narrow in scope and difficult to replicate, an issue that will be discussed in the context of other digital approaches to improving archival access.

[34] Bak, "Continuous Classification."
[35] Caswell, Cifor, and Ramirez, "'To Suddenly Discover Yourself Existing.'"
[36] Yaco et al., "A Web-Based Community-Building Archives Project."

*Knowledge Organization and the Democratizing Potential of the Digital*

As defined by Bowker and Starr, a classification system is a "set of boxes (metaphorical or literal) into which things can be put to then do some kind of work--bureaucratic or knowledge production."[37] Classification systems attempts to sort the complex realities of the world into a neat set of abstract categories, in order to achieve a specific purpose, whether administrative or conceptual. Knowledge organization refers to classification systems for information. Archival processes, such as arrangement and description, can be thought of as a form of knowledge organization for records. As discussed previously, though many archivists view archival processes as simply preserving the existing state of the records rather than as an intentional intervention, classification systems always have a purpose and so archival arrangement and description can never be truly neutral.[38] The articulation of records as the natural by-products of the function of an organization and of archival description as a neutral recording of that natural form rather than a subjective intervention by the archivist, can thus be understood of as one way that archives try and document the archival bond and authenticity of the record and thus its authenticity. Even the supposedly neutrality serves a specific human purpose.

Knowledge organization in archives is expressed through the archival processes of arrangement and description, which are recorded in the form of metadata. Each *fonds* has a finding aid, with collection-level metadata that provides contextual information about the records, including how they are organized and background about their creation. As Bak explains, "Since a record is defined as a document in the context of its relationships, and since every addition of metadata documents new relationships, the addition of metadata updates the context and changes the record."[39] Because a record's context is an integral part of its meaning, the re-contextualization provided by archival arrangement and description is also fundamental to understanding a record, and the added metadata forms a new part of the record. In other words, archival metadata both serves as an access point to the record, and as a component of the record itself. Thus metadata is a tool for the knowledge organization systems of archival arrangement

---

[37] Geoffrey C. Bowker and Susan Leigh Star, *Sorting Things Out: Classification and Its Consequences*, (MIT Press, 1999): 10.
[38] Bak, "Continuous Classification": 288-289.
[39] Bak, "Continuous Classification": 299.

and description, and those knowledge organization systems are used to organize both the records themselves and the added archival metadata, which much be understood as a part of the records.

The classification systems used to arrange and describe archives attempt to be complete, consistent, and mutually exclusive[40], despite the extreme heterogeneity of archival holdings. Archival records are often unique, since the record emerges as the by-product of some other personal or organizational function -- letters, manuscripts, personal photographs, and the like. Because the context of the record's creation is considered part of the record itself, even different copies of the same memo sent to several departments can be thought of as unique records, since they fit into the function and record collection of each department in a different way. Thus each occurance of a document is its own record and reach record belongs in a specific single place within the archival hierarchy. The sheer volume of records, both conceptual and physical, leads archives to arrange materials hierarchically, with rich historical and biographical description to give context to an entire collection and explain the relationships between its various parts. Individual items rarely have detailed description because it would be impossibly labor intensive for archivists to read and describe each item and also because archivists view that type of interpretation work to be the job of the researcher.

Common archival standards for description such as Describing Archives: a Content Standard (DACS)[41], have minimum descriptive fields needed to be able to differentiate archival collections from each other, but tend to allow extreme flexibility and deference to local convention for how to format the information that goes in each descriptive field -- for example, the extent of a digital collection could be given as the amount of storage space needed for the file in gigabytes or could be the amount of shelf-space taken up by the external hard drives or CD-ROM's that hold the materials. Additionally, because of the uniqueness of holdings and donors who are not necessarily published or conventionally famous, while archives do often use naming standards from libraries, such as the Library of Congress Name Authority File or Library of Congress Subject Headings, they also rely heavily on local rules or institution-specific

---

[40] Bowker and Star, *Sorting Things Out: 10-11.*

[41] Society of American Archivists, "Statement of Principles," *Describing Archives: A Content Standard, Second Edition (DACS)*, https://www2.archivists.org/standards/DACS/statement_of_principles.

standards. Consequently, system-wide analysis of racial[42], sexual[43], or other political[44] classification or updates to more culturally sensitive terms like the Xwi7xwa Library's First Nations House of Learning (FNHL) thesaurus approved by the Library of Congress MARC standards office[45] have occured in libraries, but not in archives.

However, in the last two decades, archives have followed the shift in libraries to digitize materials in order to improve access for users who may not be able to physically visit the holding institution in person. Digital platforms do not simply reflect a change in the location and format of the materials, but also a shift to a more user-centered approach, with a digital library that is "active rather than passive, that people would no longer go to the library, but that the library would go to the people"[46]. The digitization of materials requires a corresponding digitization of metadata, and digital metadata represents more than a difference of format but allows for new semantic possibilities. While the organization of physical records is limited by space and staff resources, digital platforms can allow for new possibilities of tailoring to individual users because they can self-customize sets of materials according to their own needs. However, findability, access, and customization for users all rely on high quality metadata, and inadequate to inaccurate metadata affects users much more greatly in a digital environment at scale where browsing and consulting an information professional are much more difficult.

Much of the thinking on the semantic potentials of digital spaces follows Weinberger's distinction of a third order of organization that can occur in digital spaces, beyond the first two that exist in and limit physical spaces. He argues that the first two orders of organization -- the arrangement of objects and the arrangement of pointers to objects are politicized because there will always be a finite number of ways to access the information. However, the flexibility of digital assets "creates a third order of order, one in which there need to be no single winner, since in open information environments like the Web the same digital information can appear in as

---

[42] Melissa Adler, "Classification Along the Color Line: Excavating Racism in the Stacks," *Journal of Critical Library and Information Studies* 1, no. 1 (January 29, 2017).
[43] Melissa Adler, "'Let's Not Homosexualize the Library Stacks': Liberating Gays in the Library Catalog," *Journal of the History of Sexuality* 24, no. 3 (September 2015): 478–507.
[44] Melissa Adler, "The Case for Taxonomic Reparations," *Knowledge Organization* 43, no. 8 (2016): 630–40.
[45] Sandy and Bossaller, "Providing Cognitively Just Subject Access to Indigenous Knowledge through Knowledge Organization Systems," 136.
[46] Marija Dalbello, "A Phenomenological Study of an Emergent National Digital Library, Part I: Theory and Methodological Framework," *The Library Quarterly* 75, no. 4 (October 1, 2005): 401.

many arrangements as there are users who care to classify it".[47] In other words, the first order is governed by space, the second by staff time and effort, but the third is open to all users and thus allows for an infinite range of potential access points and meanings.

Bak applies Weinberger's framework to an archival setting, arguing that government records can better be accessed and used with the continuous classification of records on the item-level rather than a permanent static arrangement based on function. He suggests in contrast to collection-level description, "recordkeepers enable the same records to be represented in any number of temporary aggregations, aggregations that serve specific purposes."[48] He argues that this form of arrangement more accurately reflects how users actually interact with records and more accurately documents the archival bond of portraying records in their context of use, that "relationships exist as attributes of individual records, rather than as attributes of static aggregations of records."[49] In the government records management realm Bak describes, those particular specific purposes relate to the professional duties and functions of the individual or department accessing the records. However archival platforms intended for a wider user base, such as DPLA, those aggregations are often topical in nature, generated by archivists, educators, and other users to consolidate materials across collections on particular subjects.

Additionally, while archives often have specialized holdings with specific systems of arrangement and description that best fit the particularities of their materials, users expect a generalized platform that can display a wide variety of materials with robust search features. This disconnect between archival organization and what users are accustomed to and want in digital environments provides additional justification for Bak's push for electronic records to have robust item-level rather than collection-level metadata.

> Users also want full-text indexing of all textual content, visual indexing of images through content-based image retrieval systems (such as TinEye), and access to audio, video and multimedia content through current indexing/transcribing technologies, and through new technologies as they become available. Users are accustomed to having a full range of discovery mechanisms on the Web; few can see the logic of accessing

[47] David Weinberger, *Everything is Miscellaneous: The Power of the New Digital Disorder* (New York: Henry Holt, 2007).
[48] Bak, "Continuous Classification," 302.
[49] Ibid, 302.

> records through a single human-applied classification system such as a departmental file plan, be it functional, structural or any other variety.[50]

Various multi-institutional digitization initiatives and platforms such as the American Memory Project[51] and DPLA[52] attempt to address the desire for users to search for and access records across different archives simultaneously. DPLA is able to aggregate large amounts of material with minimal effort by providing institutions with minimum metadata specifications so that heterogeneous materials are all conformed to a single standard before ingest, so that DPLA staff do not need to create or update the description for the digital objects. This also respects the intellectual ownership and authority of the holding institution, while still making those materials open to a wider audience. However, in order to make materials from the widest range of institutions available, DPLA only displays a specified set of metadata fields present in most materials, so that format-specific metadata such as transcription for video or audio or alt text for images are not aggregated or searchable through DPLA's interface.

In their discussion of shareable archival metadata, Riley and Shepherd note that the technical requirements for digital metadata to be useful in the open shared semantic web create barriers for successful implementation in archival contexts.[53] Legacy metadata requires significant cleaning and reformatting to ensure that controlled access terms are spelled consistently and records are in formats interoperable with other systems. While a reader will likely understand two items to be about the same subject despite a missing letter, digital platforms will parse them as separate and may not return items with misspelled description in a user search. And beyond simple inconsistencies in literals, Riley and Shepherd identify a deeper challenge for successful integration of archival material into shared online contexts: platforms are usually developed with libraries in mind and take item-level records as the basic unit, while archives use hierarchical and narrative longform descriptions to give the context rather than content of records. When digital platforms can often only ingest items, but controlled access points may only be provided at the collection level, archives are faced with the decision of

---

[50] Ibid, 306.
[51] Dalbello, "A Phenomenological Study of an Emergent National Digital Library, Part I.".
[52] The DPLA Archival Description Working Group, "Aggregating and Representing Collections in the Digital Public Library of America," (November 2016), http://bit.ly/dplaCollections.
[53] Riley and Shepherd, "A Brave New World."

whether or apply subjects to all items within a collection even though it may not apply to certain items or to drop the subject access point altogether.

While digital aggregation of records promises users abilities to access and consequently to make connection across materials from separate institutional repositories, particularities of the local descriptive practices of organizations and archives that later collect that organization's records present challenges for actually making metadata compatible between collections, much less understandable to users. Additionally, aggregate platforms can also facilitate semantic harm by uncritically recreating problematic archival description in a digital setting. Tasked with creating an aggregate database to consolidate seven different church and government archives related to human rights violations against First Nations communities during Canada's residential school system, The National Research Centre for Truth and Reconciliation (NRCTR) attempts to address both the problems of unintelligibility between different archival description and arrangement systems and of using colonial nomenclature standards that First Nations peoples find inaccurate or difficult to navigate.[54] The goal of this database is to allow survivors of the residential school system, human rights activists, and researchers to more easily access materials across different institutions to gain a more complete understanding of individuals' experiences as they moved through different facilities and of the residential school system as a whole. Because the seven different archives had different minimum metadata, substantial remediation and crosswalking had to to be done in order to allow faceted searching across all archives on the same platform. Additionally, many people and places had multiple names in English, French, and various Indigenous languages. Because many children had been taken from their families at a young age, they often only knew Indigenous or colloquial names for places, many of which they only knew orally and did not have standard transliterations, which made it both challenging and especially important that there be a way to indicate the multiple names for places named within the records.

Digital initiatives often emphasize standardizing existing terms and conforming records to standardized metadata formats as a precondition of updating records with culturally responsive metadata, since improving the underlying metadata structure is necessary to accommodate more

[54] Lougheed, Moran, and Callison, "Reconciliation through Description."

complex changes. However, there have also been a few platforms that incorporate Indigenous warrant and knowledge organization systems directly into their metadata infrastructure, shaping their form and not just content. The Plateau People's Project holds a variety of materials related to various Native tribes in Western Washington state, and displays archival material both within the archival bond, showing conventional archival description written by archivists, but also in the context of Indigenous knowledge systems, with description written by tribal members explaining the significance of images and artefacts within their community context[55]. By allowing multiple sets of metadata to exist for the same record, neither classification has to be excluded or subsumed into the other, another attempt at cognitive justice. Murkutu, on the other hand, defers to Indigenous warrant around what knowledge is allowed to whom, in contrast with the general archival convention of making records publicly available. Murkutu attempts to address the semantic harm caused by the non-consensual documentation and proliferation of sacred Aboriginal practices by anthropologists and others within the Australian settler state, in which both land and culture were appropriated by the White colonial regime.[56] Through specialized metadata fields, certain records can be made available only to people of a certain age, gender, or tribal affiliation or at certain times of the year, and these restrictions are determined and enforced by members of the Indigenous community themselves rather that archivists. While these these cases of Indigenous digital platforms facilitate access while still respecting Indigenous warrant and knowledge organization, they require a lot of time and labor from the communities themselves to create and to maintain, similar to community archives, and the specificity of their format limits the ability of their materials to be ingested into aggregate systems.

[55] Kimberly Christen, "Opening Archives: Respectful Repatriation," *The American Archivist* 74, no 1 (Spring/Summer 2011): 185-210.
[56] Kimberly Christen, "Archival Challenges and Digital Solutions in Aboriginal Australia," *SAA Archaeological Record* 8, no. 2 (August 2008): 21-24.

**Problem Statement: Metadata and Semantic Harm**

In the archival field, traditional notions of provenance emerge from a specific historical, cultural, and political perspective on the nature and purpose of records, which has often excluded certain communities who are conceptualized as the subjects rather than agents of record creation and collection, as well as of knowledge production and policy making, more broadly. Because of the potential of records to serve as evidence, the erasure and misrepresentation of those communities can have wide social and political impact beyond the four walls of the archives. Archival description and arrangement should be understood as forms of knowledge organization that center context and provenance in order to represent the information contained in records, with archival metadata serving as both a tool and record of the archival processes that shape the meaning of records. This thesis thus focuses on the role archival metadata plays as both a source and potential solution to what I have termed "semantic harm." "Semantic harm" is the negative consequences of the exclusion of certain worldviews from the realm of knowledge production, which affects both current understanding and future possibilities of discovery. Three of the most common forms of semantic harm in the archives are the negative or reductive classifications of certain groups of people, usage of euphemism to obscure harm, and invisibilization of communities and issues. Because of the evidential nature of records, negative, incorrect, or absent representations of marginalized communities in the archives can lead to negative self-perception, justification for oppressive policies, inaccurate conclusions in studies and research, and more. The denial of the subjectivity of certain communities causes semantic harm not only to that community, but to everyone by limiting knowledge production overall.

This focus on archival metadata and semantic harm helps bring the social justice work done by community archives and thematic collections into dialogue with knowledge organization perspectives on the potential of digital technologies for helping to achieve cognitive justice. By expanding whose voices are centered, community archives and thematic collections address semantic harm by expanding rather than rejecting conceptualizations of provenance, through increasing the contextualization of the record within its archival bond beyond the immediate source of acquisition to the wider social and political circumstances that shaped the creation of

the record. Work on socially just classification systems and metadata schemas in the field of knowledge organization take a different approach from the increase in contextual metadata used in archives. Online, the unique capabilities of digital platforms for non-hierarchical organization, user customization, and aggregation across institutions can be leveraged to allow for a greater diversity in perspectives represented in the arrangement and description of materials. However, efforts to address marginalization and improve access in the archival science and digital knowledge organization realms generally occur completely separately. This thesis thus examines one particular thematic community collection, originally created with enriching contextual metadata and later ingested into various aggregate platforms, in order to explore the impact of these two different approaches to addressing semantic on the accessibility and intelligibility of records.

**Methodology**

In order to explore the relationship between archival metadata and semantic harm across different platforms, this thesis uses the case study methodology. A case study relies on "an in-depth, longitudinal examination of a single instance or event," as opposed to "using large samples and following a rigid protocol to examine a limited number of variables."[57] The case study here is an illustrative case study, which is a descriptive case study used to "show what a situation is like."[58] Case study was selected as the methodology for this thesis because the flexibility, adaptability, and ability to examine "multiple interrelations"[59] of this particular research method makes it most appropriate for the exploration of the complex issues underlying the research questions. Both qualitative and quantitative forms of data analysis are used in the case study.

*Case Selection*

I chose to focus on Chinese American archival materials for several reasons. There is a relatively long history of Chinese immigration to the US, so there are a wide range of materials in time period, in format or medium, and in subject matter and perspective. Chinese American archival records also illustrate many of the difficulties with socially just description and access to materials related to immigrant communities, communities of color, and marginalized communities more broadly. Archival description is especially important for records that are not in English, since those materials will need to be integrated into a predominantly English record management and display system. Processing archivists must work with subject specialists and language experts to identify and describe materials, choosing whether to translate or transcribe materials, choosing what transliteration to use when there have been different transliteration standards, and making the appropriate decision what regional and temporal form of a language appears within a particular record. Due to the age of older records, there are fewer community

---

[57] Lynn Davey, "The Application of Case Study Evaluations," *Practical Assessment, Research & Evaluation* 2, no. 9 (November 1991): 1.
[58] Davey, "The Application of Case Study Evaluations": 1.
[59] Raya Fidel, "The Case Study Method: A Case Study," *Library and Information Science Research,* no. 6 (1984): 274-288.

members that have familiarity with materials to consult. And because records are often held in larger institutions as a very small fraction of massive collections on other topics, mean that materials are often processed and described by archivists who may not be subject specialists.

In order to examine the semantic impact of technical transformations, there were digitization, platform, and format requirements that needed to be met. In order to be selected, the materials needed to be originally collected as physical records, which were then digitized. Metadata created for physical collections is created to reflect the organization and structure of physical materials in folders and boxes, and when the materials are digitized, the corresponding metadata is also translated into a digital format, but often still reflects the original arrangement and description of physical materials. In contrast, born-digital materials were not considered because they were processed, arranged, and described specifically as digital materials. The collection must also be available online on both single-institution and aggregate platforms. Institutional repositories and platforms, especially for large institutions like universities, are usually heavily optimized and customized for the particular needs and standards used by that particular institution. However, aggregate platforms must be able to ingest and display materials from a wide range of institutions, so items and metadata from contributing institutions must meet standards of interoperability in order to be ingested, requiring additional technical and format updates and optimizations. Lastly, collections need to contain archival or mixed materials. The arrangement of physical materials helps to provide context to help users understand how different types of materials fit into collections, and it is relatively easy to identify material format when looking at the physical record. A user can clearly see if an item is an original or a photocopy, how large images are, and how items are arranged in relation to each other, for example whether a photograph is part of an album or paperclipped with other materials as part of a portfolio. Displayed digitally, items rely on accurate metadata and arrangement of materials on the website to provide similar context about format and organization.

When beginning my search for Chinese American archival materials on DPLA, I quickly ran into a problem: all the items with “Chinese” in the description were of unnamed Chinese people or objects where “Chinese” was the only real descriptor, while items with detailed information about Chinese individuals, places, or cultural artefacts, had description specific to

the content of the item, such as the name, and often did not contain the additional descriptor "Chinese." Despite the relatively long history and large size of the Chinese American community, there are only two complete collections that are digitized, on institutional and aggregate platforms, and showcase a variety of item formats: the Chinese in California Virtual Collection and the Chinese Historical Society of Southern California Collection. Much of the digitized Chinese American archival materials are either small selections for class projects displayed on a blog-style website or minimally identified material from large collections such as photographs captioned "Chinese man" or "Chinese woman" within a newspaper archive. Furthermore, while the Chinese Historical Society of Southern California Collection is displayed on University of Southern California's digital library platform and on DPLA as an archival collection, an in-person visit showed that the collection is actually purely archaeological. All items within the collection are artefacts excavated from two Chinese American sites, and organized only with an inventory catalog -- no true archival organization or description exists for these materials. This leaves the Chinese in California Virtual Collection as the only option available for this case study.

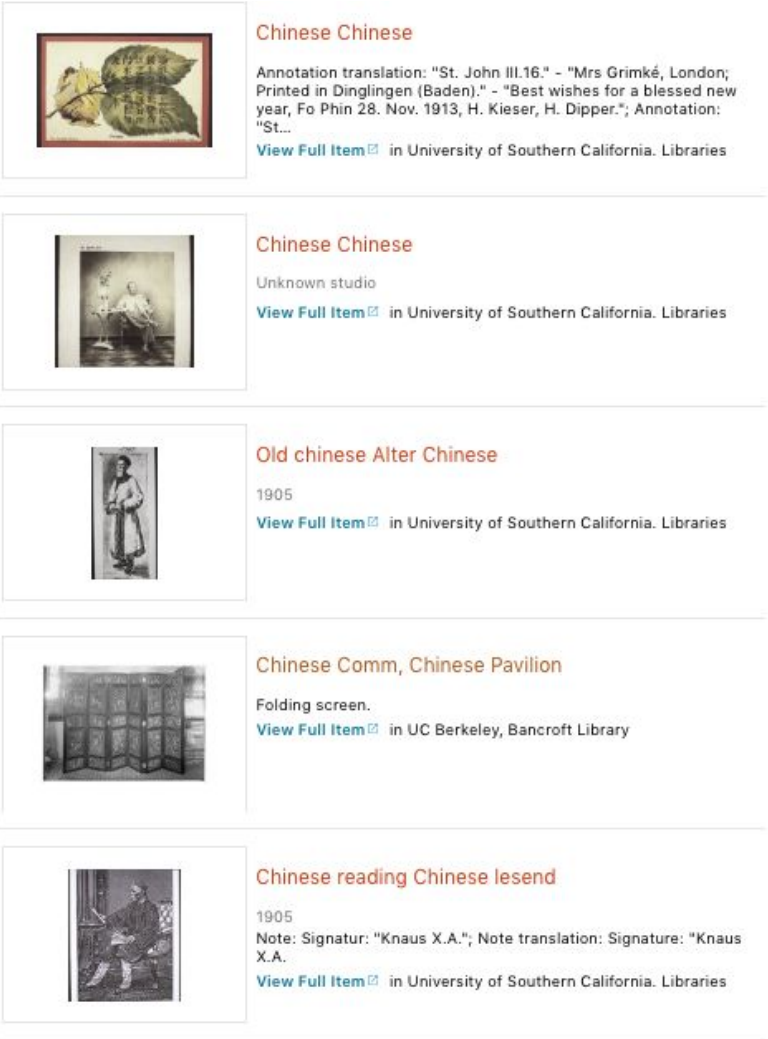



Figure 1. First five results from a DPLA search for "Chinese." Three are of unnamed Chinese people, while the other two items have equally vague titles and lack sufficient description to really make sense of the images.

The Chinese in California Virtual Collection covers materials related to the Chinese diaspora in the state of California between 1850 and 1925, drawing materials from the Bancroft Library and the Ethnic Studies Library at the University of California, Berkeley, the California Historical Society, and the Oroville Chinese Temple and Museum Complex, and was digitized by funding the Bancroft received from the 1998-1999 Library of Congress and Ameritech grant cycle. These materials are in a variety of formats ranging from photographs to postcards, from advertisements to legal summons, from paintings to artefacts, and cover a wide range of sources and perspectives, from both within and without the community. The information about the collection and the digitized items are available across four platforms: the Bancroft Library's project-specific websites, the Online Archive of California (OAC), Calisphere, and DPLA.

*Data Collection*

In order to better understand the collections, I started with in-person visits to the holding institutions. This was in order to see how the items selected for the virtual collection fit into their original (traditional single-*fonds* archival) collections, and what the process would be like for a user to find, access, and understand these materials at a physical repository. In particular, I kept notes on the ease of visiting the archives, of requesting and viewing materials in the reading room, and of making sense of the organization and meaning of records. There was not much information anywhere on the selection criteria for items within the *Chinese in California* virtual collection, so I also wanted to see the records in context to see if there was some pattern to why they were chosen and what other adjacent or similar materials were not selected for digitization.

I also did a thorough web search for all websites where information about the collection and digital records were displayed, since clicking through the various websites produced only partial linkages between the platforms. This yielded the four platforms listed previously. The different purposes of these platforms was gathered from website documentation and in-person and email interviews with staff members. The full finding aid was downloaded from OAC in both PDF and HMTL formats for manual and programmatic analysis. The metadata records for the collection were scraped from Calisphere and DPLA using Python, and converted into tables

to aid quantitative analysis. Screenshots were taken of all websites to provide case examples of the same image across platforms, and to capture the overall layout of the each platform.

*Data Analysis*

A combination of qualitative and quantitative analysis was applied to each platform and the relations between them. Collection-level and item-level metadata are analyzed on each platform, as well as the linkages between the websites. Collection-level metadata is examined in order to determine how the collection is placed within archival context for the user. First, I examined the format the collection overview or landing page takes, what access points like search and filter options exist to help the user navigate the collection, and whether or not platforms with less contextual information provide links for users to learn more about the collection as a whole. Then I examined what metadata fields or other descriptive information exist for the collection, with a special focus on what types of contextual information were available: whether there is information 1) about the contents of the collection or the historical context of the records, 2) about the creator or the context of record creation, and 3) about the archival process or the archival context that determined material selection and description. Lastly, I examined at how the items appear in relation to the collection, and whether or not there is information provided to place the record within its archival bond: how are items incorporated into the collection-level description and the platform and what information about the items is available from the collection view.

I then moved on to an analysis of item-level metadata for the each platform. When a user chooses to view a single item, how does is the item displayed to the user? I examined how views of each item is available, along with what different metadata fields exist in each view. In particular, attention is paid to whether or not items have information about how they fit into the collection as a whole, whether subject access point or scope and context notes are present, and how easily the user can find their way back to information about the collection. An overview of the items in the collection shows what metadata fields exist for items on each platform, and what proportion of the items have values for each metadata field. An example of one particular item

across all platforms was used to illustrate and explore the semantic implications of the different metadata fields between the different platforms.

Lastly, I map the relationship between the websites, since the platforms have different information and are linked to each other in different ways. Diagramming how the item-level and collection-level pages connect to each other within and across platforms shows how the different purpose and design of each platform shapes how materials are meant to be understood and interacted there, and how users are directed to other places for other types of information, even though all four platforms are displaying the same collection. Furthermore, the directionality of the linkages -- whether a user can get from collection-level to item-level display as a easily as the reverse -- and number of clicks between different types of information helps explore what information users are likely to get access to depending on how and on what platform they first come across the materials in this archival collection.

**Findings**

The four platforms that display the *Chinese in California* virtual collection are the Bancroft Library's project-specific websites, OAC, Calisphere, and DPLA. The platforms are arranged in order from oldest to newest, which also reflects increasing scope of aggregation. The Bancroft's original project site gives the historic context for the creation of the collection and for the history of the *Chinese in California*, as well as detailing the technical and archival processes undertaken to make the project a reality. It is the oldest platform and contains no aggregation at all. OAC hosts the full finding aid for the collection, as well as the items themselves, while Calisphere and DPLA are geared primarily towards hosting the digital objects themselves, though the more item-based platforms do link back to the ones with collection-level description. OAC and Calisphere are both aggregate platforms designed by California Digital Library, a branch of the University of California System; OAC aggregates finding aids and older digital records statewide, while Calisphere hosts and aggregates all types of digital objects for the state of California, beyond the archival scope of OAC. DPLA is the newest platform, and aggregates archival, bibliographic, and other material all across the United States. The Bancroft's websites and OAC are designed specifically for archives and allow for rich descriptive metadata as way to increase the archival and historical contextualization and thus enrich the meaning of records. Calisphere and DPLA are designed to display individual items of all types and allow access to a wider range of materials all in one place by freeing items of their specific context through aggregate. Platforms with collection-level information sought to describe the records within their historical and archival context, with rich prose description to explain how the record came to be and how it came to be part of the collection, while the item-based platforms allow the user to access materials from a wider range of collections and institutions and look more modern, but are missing much of the metadata and context and would allow the item to be found or understood.

*Bancroft Original Project Site*

The Bancroft Library's project-specific websites provide the most archival context out of all the platforms, while also providing fairly robust historical context for understanding the content of the collection. It does not format information in the conventional archival metadata

format of a finding aid, but does provide rich prose detail about the history of and archival processes undertaken to create the collection. Information about how the project was funded, where it was hosted online previously, the technical specifications digitization and content management, and the underlying rationale for the arrangement and description of the collection can all be found on the project-specific website, and not in the finding aid itself, which is hosted on OAC. Additionally, though the digitization and processing of materials from the Oroville Chinese Temple was funded through and undertaken as part of the *Chinese in California* project and those records added as a series within the collection, the Oroville Chinese Temple sub-project has its own separate website[60]. The Oroville Chinese Temple sub-project website contains information about the temple's history and technical specifications for photographing artefacts and digitizing images, as well as its own unique search interface that allows browsing for each room of the temple, but can only be reached through one particular page on the main *Chinese in California* project website. Despite the rich historical and archival context created for the sub-project and provided through its website, the information about the Oroville Chinese Temple series was not incorporated into the finding aid and is extremely difficult to find, since it is only linked from one place.

The emphasis on making visible the role of the various archivists and archival institutions can be understood as enriching the provenance of the collection by acknowledging how archival processes interact with historic social and political contexts to shape records, especially about communities who are the subjects of records, such as Chinese Americans. In particular, through the discussion of materials selection and arrangement and description choices[61], the Bancroft original project site touches on semantic harm done to Chinese Californians and how this project sought to address or correct those semantic harms. For example, the Bancroft explains that "[b]ecause of the complexity of the social and political history of the time, the presentation of the materials is organized thematically. In selecting the material and constructing this digital archive it became apparent that much of the material reflected an outsiders' view of the Chinese communities." The Bancroft acknowledges that many of the records about Chinese Americans

---

[60] Bancroft Library, "Oroville Chinese Temple," (December 11, 2006), http://bancroft.berkeley.edu/collections/oroville/.

[61] Bancroft Library, "About the Project," *Chinese in California, 1850-1925* (December 3, 2008), http://bancroft.berkeley.edu/collections/chineseinca/about.html.

are not created or collected by them, so they choose to organize series by topic instead of relying on a *fonds*- or original order-based approach, which would center the largely non-Chinese perspectives of the creators or collectors of records. They note a general scarcity of material in the archives, especially from the community itself, though they acknowledge that the Ethnic Studies Library has much more primary source material, implicitly pointing out that the goals and interests of an archival institution shape what subjects are collected and made present or absent in the historical record. They also acknowledge that many records containing "caricatures and derogatory designations" were selected for this collection and are used for research because of the relative scarcity of materials. This demonstrates an awareness of how the semantic harm of misrepresentation impacts how new knowledge can be formed, as well as a commitment to confronting and addressing that misrepresentation by providing additional historic and archival context for those records.

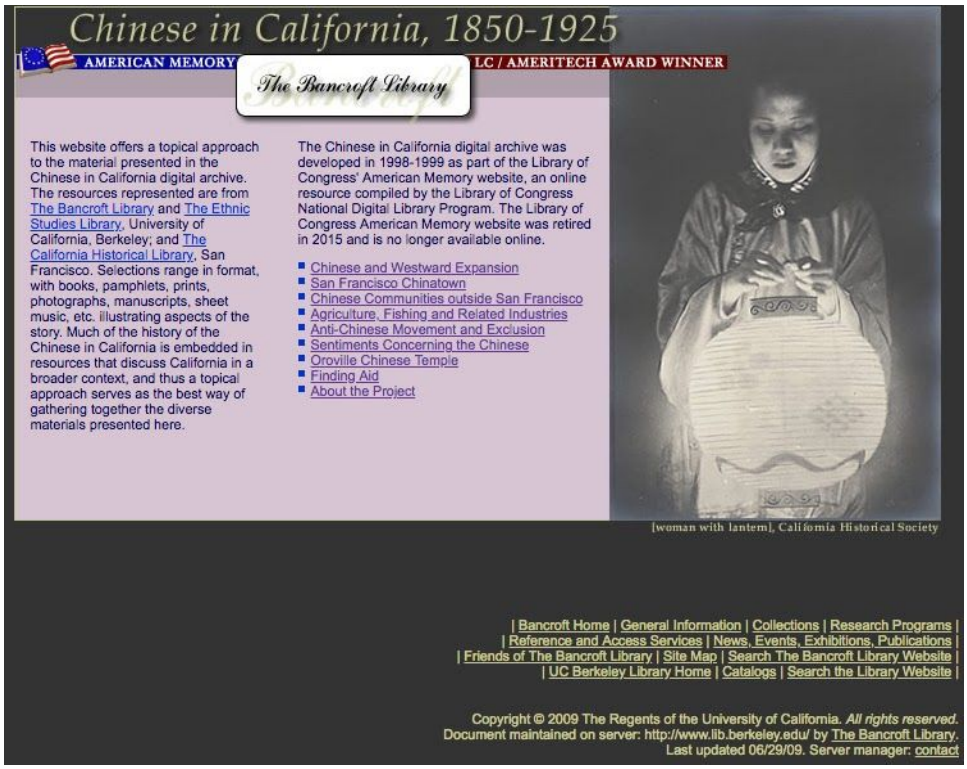



Figure 3: Bancroft original project site home page. The "Finding Aid" link takes the user to the full Finding Aid on the OAC website, while the other links take the user to pages about the different series within the collection, as well as the technical specifications and archival processing background for the project.

Because of how many records were outsider accounts or images about rather than from the Chinese community, series are determined by topic, and description in the finding aid is provided at the collection and item levels. The Bancroft original project site does not actually list

the items in the collection or host the finding aid, but explains the technical and organizational principles behind the creation of the finding aid with all items embedded within. Organizing materials by topic is not an abandonment of provenance, but rather an expansion of provenance beyond the immediate source of acquisition of the materials by providing context for the materials by highlighting the broader social, political, and cultural dynamics that lead to the scenes and events depicted in the records and to creation of these materials.[62] The series in the collection reflect broad historical trends, such as "Westward Expansion," which portrays the early waves of Chinese immigrants to arrived during the gold rush and the building of the transcontinental railroad, and "Anti-Chinese Movement and Exclusion," which engages actively with the discriminatory social and political mobilization that arose around anti-Chinese sentiment. The importance of collection-level description to show the record within its historic and archival context is emphasized:

> The data represented a mix of item-level cataloging and collection-level cataloging. Collection-level cataloging is useful for presenting users with a contextual view of an archival collection. Showing items in the context of its series and subseries and neighboring items. An item-level view, such as an interface to browse, search, and sort individual images, necessarily presents items out of context.[63]

The statement that item-level views present records out of their context will be revisited in the subsequent platforms, though it is interesting to note that the creators of the project were aware of the loss of context that would occur in a platform that only allows for item- and not collection-level description and organization. The active engagement with and depiction of archival processes as part of the record collection itself aligns well with the record continuum model, where archivists preserve and enrich the meaning of records once they have been removed from their immediate context of creation and use into the archive through additional layers of metadata provide broader (re-)contextualization.[64]

---

[62] Millar, "The Death of the Fonds and the Resurrection of Provenance": 12-13; Douglas, "Origins": 24.
[63] Bancroft Library, "About the Project."
[64] Upward, "Modelling the Continuum as Paradigm Shift in Recordkeeping and Archiving Processes and Beyond"; McKemmish, "Placing Records Continuum Theory and Practice": 335-336.

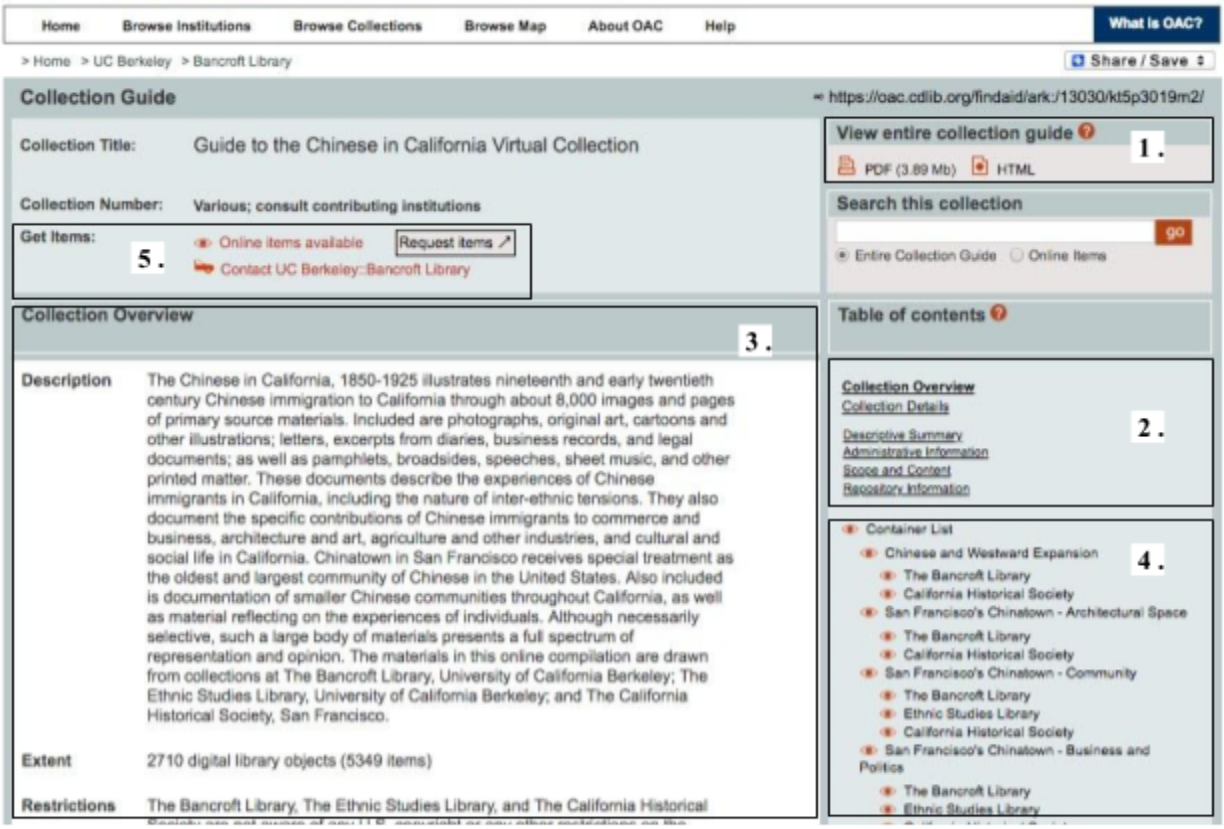



Figure 3. Collection overview page from the finding aid on OAC. 1. Download links for the finding aid in either PDF or HTML format. 2. Navigation to the collection overview (currently shown) or the collection details. 3. The part of the finding aid the user is currently browsing shows up here. 4. Click "Container List" or a particular series or sub-series to jump to that portion of the finding aid, and view series- or source collection-level metadata or access individual items. 5. Items can also be viewed by clicking the eye next to "Online items available."

*Online Archive of California (OAC)*

OAC is a platform that hosts finding aids for the University of California system, as well as other archival institutions throughout the state, including the finding aid or guide to the *Chinese in California* collection. The OAC landing page[65] provides an overview of the collection, while the collection detail page[66] (figure 3, item 2) provides more in-depth information about the three original contributing institutions -- the Bancroft Library, the Ethnic

[65] Online Archive of California, "Collection Overview," *Guide to the Chinese in California Virtual Collection*, https://oac.cdlib.org/findaid/ark:/13030/kt5p3019m2/.
[66] Online Archive of California, "Collection Details," *Guide to the Chinese in California Virtual Collection*, https://oac.cdlib.org/findaid/ark:/13030/kt5p3019m2/admin/.

Studies Library, and the California Historical Society, administrative information about the use of materials such as publication, citation, and access, and a general overview of the structure of the collection. Information about the funding of the project, record and metadata digitization process, and archival arrangement and description choices are not present, and there is no link back to the Bancroft original project site where that information or any of the information about the Oroville Chinese Temple series can be found. Users can download a full version of the finding aid in either PDF or HTML format (figure 3, item 1), and the items can be viewed through either the finding aid container list[67] (figure 3, item 4) or through a gallery[68] accessed by clicking the "Online items available" link in the "Get Items" section (figure 3, item 5).


Figure 4. Side by side comparison of item access through the gallery (left) and finding aid (right). In the gallery view, thumbnails are nested under the series, sub-series (contributing institution), and source collection. In the finding aid, links to specific items follow a similar arrangement, though more contextual metadata is provided. Just prior to the section capture here is a paragraphs-long prose description of the Chinese and Westward Expansion series that these two items are a part of. They are shown here listed under the appropriate sub-series (materials sourced from The Bancroft Library) and the source collection (Photographs from the Burckhalter Family collection), along with a source collection scope and content note and subject and indexing terms.

The two ways of accessing items on OAC, through the finding aid directly and through the gallery, which is also hosted directly through OAC. Both the finding aid and gallery list the series, contribution institution, and source collection, and then list by title all the items from the same source collection together, which reaffirms the archival bonds between records within and with their source collection. Because the items in the *Chinese in California* collection share the

[67] Online Archive of California, "Container List," *Guide to the Chinese in California Virtual Collection,* https://oac.cdlib.org/findaid/ark:/13030/kt5p3019m2/dsc/#dsc-1.8.5.
[68] Online Archive of California, "Online Items," *Guide to the Chinese in California Virtual Collection*, https://oac.cdlib.org/view?docId=kt5p3019m2;developer=local;style=oac4;doc.view=items.

same topic but are drawn from different archival collections, source collection refers to the original archival collection the item is a part of. In addition, in the finding aid, each series has a scope and content note that provides about a page of detailed historical context for the items in that series. Even for items that do not have source collection- or item-level scope and content notes or subjects, seeing the item within the context of the title, creator, and date of the source collection and among the other items from the source collection can provide valuable insight into understanding the item. Scope and content notes or subject access points may also be provided for the source collection or the individual item. The gallery view shows only image thumbnails and titles and does not list any other information about the series, source collection, or item.

When examining the archival metadata for the 2,710 items within the *Chinese in California* collection, three primary factors are taken into account as portraying the provenance and context of the record: source collection type, the presence of descriptive metadata or subject access points, and the level of those metadata fields. Source collections are broken into two types: archival collections and subject files. Archival collections in the traditional sense comprise a single *fonds* arranged in original order, though here the criteria are slightly less strict -- collections created or collected by a known individual or organization source are considered archival. Subject files consist of collections of loose materials organized by subject rather than source, and the creator or collector is often unknown. About three fourths of the items in the collection come from traditional archival collections, while the remaining fourth is drawn from subject files. The archival source collections are generally arranged and described following professional archival conventions, and often have their own finding aids to draw information from, though the finding aid of the original source collection is not linked from the *Chinese in California* finding aid.

Two metadata fields were chosen to represent context: “scope and content note” and “subject” (which occasionally also appears as “subject and indexing terms”). The scope and content note generally gives an overview of the source collection or item, providing information about the context of the record’s creation, while the subject uses Library of Congress subject headings to describe what the record is about, including the cultural and historical context portrayed within the record. A scope and content note serves as an active attempt to record and

report the detailed circumstances surrounding the creation of the record, while a subject helps establish the broader social and political context of the record. These two metadata fields can be present at the source collection- or item-level, both, or neither. Generally source collection-level metadata is longer and provides more detailed context, as seen in figure 4, while item-level metadata provides a shorter, simpler description of the item itself. Items from subject files are more likely to have item- rather than source collection-level description, since those records do not belong to any identifiable *fonds* and thus are more difficult to place in their archival context. They are also more likely to have a subject rather than scope and content note, which at least provides broad historical context for the record. All items from the Ethnic Studies Library have only source collection- and not item-level metadata. Because of the Ethnic Studies Library's mission to preserve and highlight the voices of marginalized communities of color, they have more collections directly created by Chinese American communities, which is noted in the About page of the Bancroft's project-specific site. The prevalence of materials from archival collections rather than subject files combined with their semantic justice-oriented vision allow the Ethnic Studies Library to provide the richest possible contextual information, which occurs at the source collection-level to emphasize the historical context and archival bonds of each record. In contrast, none of the source collection-level information for the Oroville Chinese Temple was transferred from the project-specific website to the finding aid. Those items have the sparsest


AAS ARC 2000/43 **Chinese League of Justice of America miscellany 1909-1910**

**Creator/Collector:** Worley, Franklin B.

**Scope and Content Note**

Contains typescript reports of July 26 and Oct. 8, 1909 made by the League's attorney, Franklin B. Worley, regarding his activities in response to negative publicity against the Chinese, particularly concerning the Chinese mission schools and the reported murder of Elsie Siegel by Leon Ling. Includes Bulletin No. 1, issued by the League on Sept. 1, 1909, devoted to a discussion of statements made by Sir Robert Hart, for many years an official in China, concerning Chinese converting to Christianity. Also includes a letter, dated Aug. 20, 1910, from Charles S. Whitman, District Attorney for New York county, enclosing a copy of a report from the Police Commissioner in response to complaints against the showing of the picture, "The Chinatown Trunk Mystery," based on the Elsie Siegel murder.

**Subjects and Indexing Terms**

Chinese -- Missions -- California -- San Francisco.
Chinese Americans -- California -- San Francisco -- Social conditions.
Chinese Americans -- California -- San Francisco.
Chinese Americans -- Civil rights -- California -- San Francisco.
Chinese Americans -- New York (State) -- New York.
Discrimination against Chinese.
Discrimination in criminal justice administration.
Hart, Robert, -- Sir, -- 1835-1911.
Siegel, Elsie.

AAS ARC 2000/43: 26 July **"Report of Progress Made So Far" 26 July 1909**
ark:/13030/hb8n39n9bz

**Contributing Institution:** Ethnic Studies Library

Figure 5. Example of an item contributed from the Ethnic Studies Library, with scope and content note and subject access points listed only at the source collection level. These provide details about the creation and contents of the source collection that help contextualize the individual items listed below, which only have a call number, title, and link.

archival and historical context -- the lack of even series-level description means there is almost no explicit context for the materials provided beyond a contributing institution and a title.

Approximately one fourth of items do not have a subject or scope and content note at either source collection- or item-, though over half of those are the items from the Oroville Chinese Temple. Though some basic context is implied for those items through the name of the series they are contained in, the lack of description or subject access point makes these items more difficult to find or understand. Additionally, about one quarter of items from archival collections and one fifth of items from subject files only have scope and content notes and subjects at the source collection-level, including all items contributed by the Ethnic Studies Library. Because only item-level metadata is ingested by Calisphere and DPLA, those items will suffer an almost complete loss of contextual information on those platforms.

| | | **Subject** | | | |
|---|---|---|---|---|---|
| | **Scope and Content** | Both | Source Collection | Item | None |
| **Archival Collection (n=2041)** | Both | 14 (1%) | 0 | 294 (14%) | 0 |
| | Source Collection | 103 (5%) | 491 (24%) | 228 (11%) | 19 (1%) |
| | Item | 0 | 0 | 52 (3%) | 0 |
| | None | 0 | 5 (<1%) | 151 (7%) | 684 (34%) |
| **Subject Files (n=669)** | Both | 11 (2%) | 9 (1%) | 0 | 0 |
| | Source Collection | 0 | 130 (19%) | 10 (1%) | 0 |
| | Item | 0 | 0 | 217 (32%) | 0 |
| | None | 0 | 12 (2%) | 280 (42%) | 0 |

Table 1. Breakdown of archival metadata related to context for all items. Items are first divided between those from archival collections versus those from subject files, and then broken down further between those that have scope and content notes and/or subjects at the source collection-level, item-level, both, or neither. For example, in the second cell of the first column, 103 items have subject access points listed at both the source collection- and item-level, but only have a scope and content note present at the source collection- and not the item-level.


Figure 6. Side by side view of the OAC image view (left) and the OAC image detail page (right). The image view contains a large image, along with the title, contributing institution, and a link to a more detailed description. The image detail page contains a small thumbnail of the image accompanied all the metadata that exists on the item-level. The Related Item field also shows where the item fits into the larger collection, listing first the series, then the sub-series (contributing institution), and then finally the source collection.

There are two different item-level view pages, an image view with just the item and title and a detail view with a thumbnail and full item-level metadata. For single-image items, the user is taken first to the image view with a link to the detail view, while for multi-image items like books, the user is taken first to the detail view with a link to view all pages in the item-view. It is not clear from the finding aid which items have multiple pages, or even any multi-page items are part of the collection. In the image detail view, all items have a title, identifier (the call number for the item in the physical repository), contributing institution, and related item field, which somewhat counterintuitively states the series and source collection rather than listing other related items. There is also a link back to the OAC overview page in the collection field. If item-level metadata exists for scope and content note, subject, creator/collector, alternate title, or physical description field, they will also appear in the image detail view. Only item-level metadata is pulled into the image detail view; source-collection metadata is not inherited. This is important to note, as the more item-focused platforms, Calisphere and DPLA, only ingest item-level metadata from OAC, and do not provide access to source collection-level metadata.

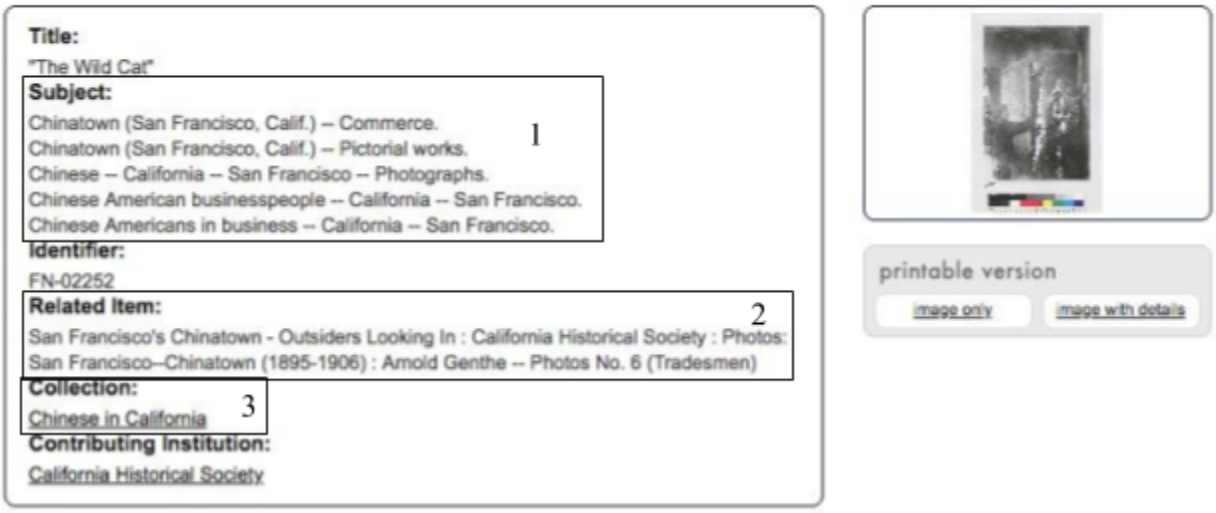



Figure 7. The image detail view in OAC for "The Wild Cat." 1) Subject access points. 2) Related item field showing the series (San Francisco's Chinatown - Outsiders Looking In), the sub-series (California historical Society), the source collection (Arnold Genthe's photographs of San Francisco Chinatown), and the series within the source collection (Tradesmen). 3) Link back to the finding aid for the *Chinese in California* collection.

To look at an example, let us examine the OAC image detail page[69] for "The Wild Cat"[70]. From the related item field (figure 7, item 2), we can see that this particular item is from California Historical Society contribution to the series "San Francisco Chinatown - Outsiders Looking In." In particular, it was selected from the "Photos No. 6 (Tradesmen)" series of the source collection *Photos: San Francisco Chinatown (1895-1906) by Arnold Genthe* source collection. From the series scope and content note provided in the finding aid series, we can ascertain that the "San Francisco Chinatown - Outsiders Looking In" series was created to highlight the American fascination and exotification towards the Chinese community, which provides the context that records within this series are either created by community outsiders or by Chinese people engaging with tourists. This thematic series organization complexifies archival provenance beyond the individual who created or collected these items by contextualizing them within broader sociocultural dynamics of the orientalism of that time period, naming and making known the white gaze as specific and culturally informed, a stark and intentional contrast to how archival records often appear as objective evidence of past events. A

[69] Online Archive of California, "The Wild Cat," *Chinese in California*, https://oac.cdlib.org/ark:/13030/hb3w1002x7/?brand=oac4.
[70] Arnold Genthe, "The Wild Cat," Photos: San Francisco--Chinatown (1895-1906): Arnold Genthe -- Photos No. 6 (Tradesmen): FN-02252, California Historical Society, San Francisco.

link back to the finding aid (figure 3, item 1) allows a user to gain additional information about the collection, if they are accessing the item directly. The collection source series title "Photos No. 6 (Tradesmen)" supplements the subjects related to business and businesspeople (figure 7, item 1) to portray the subject of the photograph as someone who is in the business of buying or selling wild cats or their products. Though no source collection-level subject or scope and content note is provided for this source collection, we can see from the title, other series, and other items within this source collection, when Genthe took his photographs and what types of records he created. For example, from the first series within this source collection, "Photos No. 1 (Camera Shy Chinese)," items such as "Fleeing From the Camera," "An Unsuspecting Victim," and "No Lickee" -- a photo of a Chinese man covering his face, where "lickee" is either a mockery of a Chinese pronunciation of "like" or "look," confirm Genthe to be a community outsider who was not particularly concerned with ethics of getting consent from photographic subjects. Though these three records appear somewhat offensive and degrading, the additional framing done by the series-level description allows users to be prepared to encounter this type of material and to better understand how records like these came to be created and deposited into archives, as well as their impact on Chinese American communities and on research and scholarship done on this particular community.

*Calisphere*

Though OAC and Calisphere are both products of CDL, a branch of the University of California system, and share similar partner institutions, Calisphere is newer than OAC and is designed for the direct hosting and display of items. While Calisphere does host plenty of archival materials, among other types, it is not designed for collection-level archival description or arrangement, such as a finding aid. The *Chinese in California* collection is displayed as four separate collections within Calisphere, one for each contributing institution. A shortened version of the collection abstract from the finding aid appears at the top of each gallery page (figure 8, item 1). However, the shortened abstract does not name the different contributing institutions or the galleries do not link to each other, so there is no way to know that each digital collection is just a portion of *Chinese in California*. Aside from the abstract, there is no additional

information about the collection, although there is a link to the full finding aid on OAC (figure 8, item 3). The Bancroft's gallery page[71] also links back to the Bancroft original project site (figure 8, item 2), the only page that does so across all platforms. Unlike the OAC gallery, which displays items by series, institution, and source collection (figure 4), the Calisphere gallery by default displays items in alphabetical order (figure 8, item 6). OAC did not allow for any filtering from the gallery or finding aid, but on Calisphere, items can be filtered by type -- whether the item is an "image" or "text" -- or by decade (figure 8, items 4 and 5). Other metadata fields are not filterable, although a string-match search can be performed with the search box. In contrast to the rigid pre-determined arrangement of the items in OAC, the flexibility of the Calisphere gallery to display items in different ways reflects the potential of the digital to allow users to

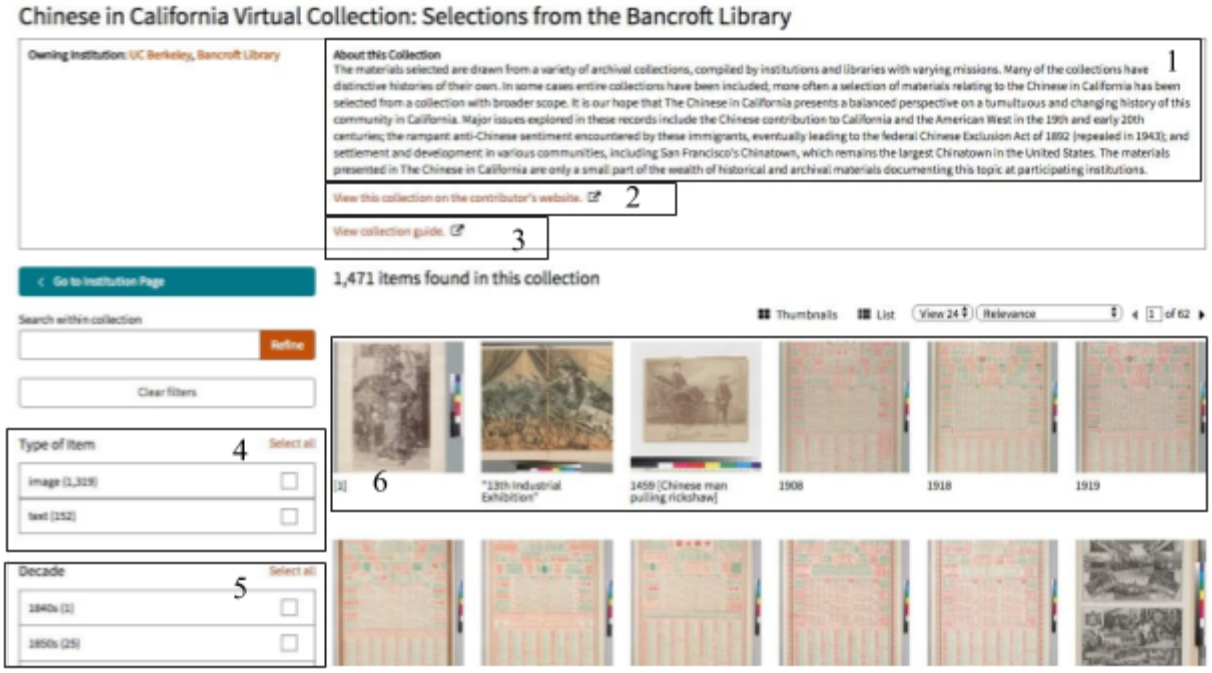



Figure 8. Calisphere gallery page for *Chinese in California* items selected from the Bancroft Library. Though the gallery does specific the contributing institution, there is no indication that this collection contains selections from four separate institutions and there is no link between this gallery and the galleries for the other three contributing institutions. 1) A shortened version of the abstract giving an overview of the collection. 2) A link to the Bancroft's project-specific website. None of the other Calisphere galleries or any other pages on any other platforms link back to the unique and valuable archival and historical context provided there. 3) A link to the full finding aid on OAC, described here as a collection guide. 4) Option to filter by type of item: image or text. 5) Option to filter by decade. 6) The images themselves are displayed as thumbnails with titles. The default organization is alphabetical order by title, a stark contrast to the series, subseries, and source collection organization presented in OAC (see figure 4).

[71] Calisphere, "Chinese in California Virtual Collection: Selections from the Bancroft Library," https://calisphere.org/collections/143/.

access and organize information in their own ways[72]. Likewise, the wide range of materials, archival or not, available in a single consolidated place through Calisphere, allows users to more easily access a wider range of topics and media in a single place. This is especially critical for improving the reach of archival records, since many people are not familiar with archival records and may not think to seek them out specifically.

Clicking on an image thumbnail or title takes the user to a separate page for that particular item. While it initially appears that the Calisphere item page has more metadata fields than the OAC item detail view, the additional fields are modified standard statements added to all items in the collection, such as publication information -- the address and website of the contributing institution, collection -- a link back to the gallery page (Figure 9, item 2), and a rights statement. The Calisphere item page shares the same title and identifier as the OAC item-detail view, and includes all optional item-level metadata fields, albeit with some format modification. The OAC subject gets pulled directly, unless it lists a geographic location as its first term, in which case that subject moves into a separate "place" field. For complex subjects delimited with a double dash, each item was separated into its own place (figure 9, item 3). For example, the OAC subject "Chinatown (San Francisco, Calif.) -- Commerce" becomes two separate Calisphere places, "Chinatown (San Francisco, Calif.)" and "Commerce." The scope and content note is ingested directly word-for-word but is renamed "description," which is assumed to be more user-friendly than the very archives-specific "scope and content note". If a creator/contributor is listed -- always a single field in OAC, separate creator and contributor fields are lists, though the two contain the same information.

The information in the "related items" field, which lists the source collection in OAC, is completely missing from Calisphere. Because Calisphere allows users to search for items across many collections and institutions, the source collection information may appear too context-specific and unnecessary for the user to find and access a record. There are links back to the OAC item page right under the image or in the first line of the metadata section (figure 9, item 1), so the user can also navigate to that information if they need to. Single-page items link back to the OAC item page, while multi-page items link back to the OAC item detail page. Like

[72] Weinberger, *Everything is Miscellaneous*; Bak, "Continuous Classification."

in the OAC finding aid, there is no indication in the Calisphere gallery or item view that some items have multiple pages.

Like the OAC image views, the Calisphere image view only lists item-level metadata, with no inherited series- or source-collection level information. However, by omitting the series and source collection, it becomes much more difficult for the user to find additional context for the record through the Calisphere item record than through the OAC item detail view. Because Calisphere only supports item- and not source collection-level metadata, nearly half of the items now have no subject or description, double the number of items in OAC with no contextual metadata at all. While items without a subject or description can be viewed when browsing through Calisphere's galleries, they are extremely difficult to search for, since the only unique information about those items are contained in the title -- most of the metadata fields like the rights statement or publication information are generalized for the whole collection and not particularly useful to identifying any one particular record. Furthermore, for items that are part of an archival collection, the title may only make sense in the context of viewing the item together with the other items with the source collection-level metadata. For example, the first three pages of one photo album[73] are titled "[cover]", "title page," and "[1]", and have no item-level scope and content note or subject, though both exist at the source collection-level in the finding aid. Because all items from the

| | | Subject | No Subject |
|---|---|---|---|
| The Bancroft Library (n=1471) | Scope | 343 (23%) | 12 (1%) |
| | No Scope | 408 (28%) | 708 (48%) |
| Ethnic Studies Library (n=240) | Scope | 0 | 0 |
| | No Scope | 0 | 240 (100%) |
| California Historical Society (n=634) | Scope | 143 (23%) | 14 (2%) |
| | No Scope | 447 (71%) | 30 (5%) |
| Oroville Chinese Temple (n=365) | Scope | 0 | 2 (1%) |
| | No Scope | 0 | 363 (99%) |

Table 2. Breakdown of how many items in each of the four separate Calisphere galleries for each contributing institution have contextual metadata about the record's creation ("scope and content note" pulled from OAC, renamed "description" in Calisphere) or the subject matter portrayed in the record (subject pulled from OAC, separated into "subject" and "place" in Calisphere; both "subject" and "place" are considered part of "subject" for the purpose of this table). Particularly noteworthy is that because the Ethnic Studies Library provided scope and content note and subjects at the source collection level for all items in the OAC finding aid, all the information has been lost in the move to Calisphere.

[73] Isaiah West Taber, *Souvenir of California: photographic views San Francisco, Cal.* (Taber Photo., [ca.1882]: pfF869.S3.9.S718), The Bancroft Library, University of California, Berkeley.

Ethnic Studies Library have only collection-level subjects and scope and content notes, in Calisphere, now none of them have any subjects or descriptions. Additionally, even for items that in OAC have both source collection- and item-level metadata, the loss of source collection-level information removes a lot of context from the record.

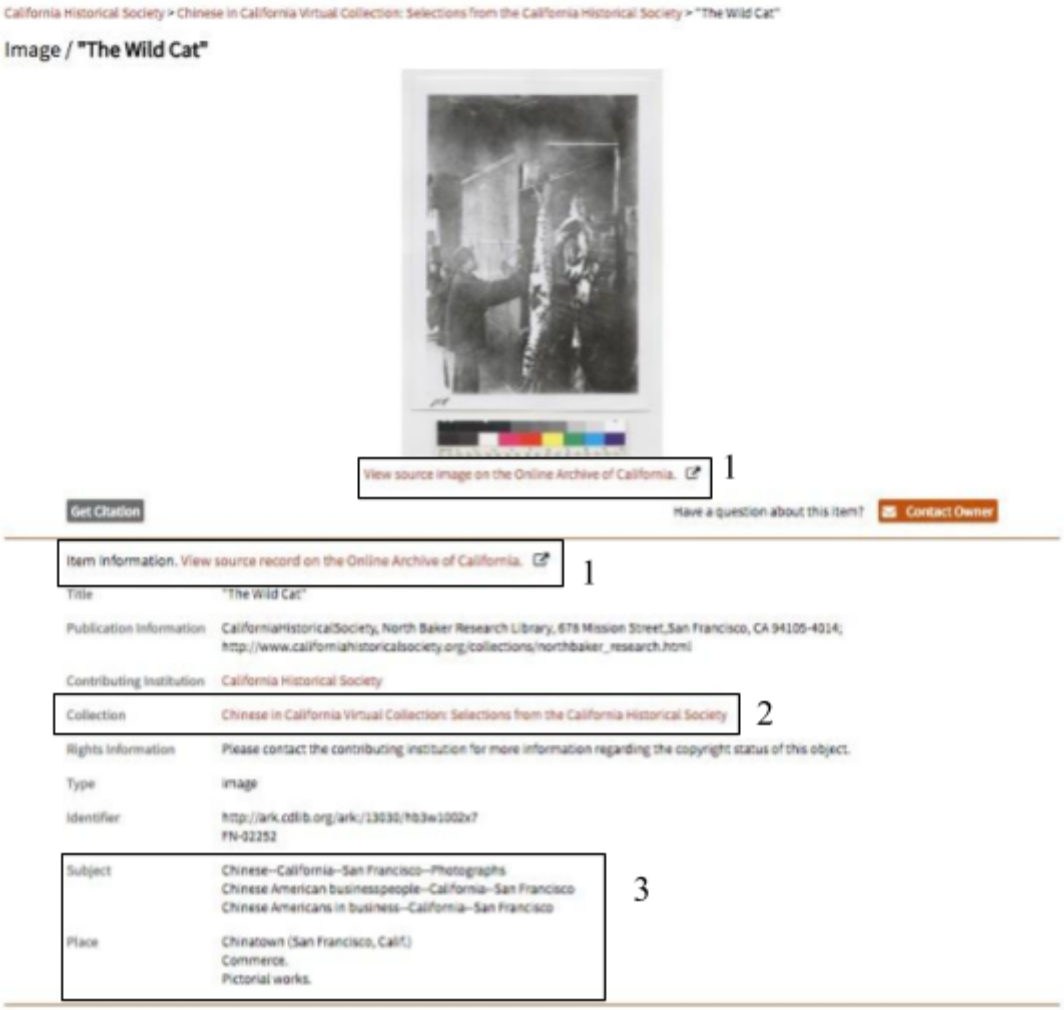



Figure 9. Calisphere image page for "The Wild Cat." 1) Links to the OAC image detail view. 2) Link back to the particular contributing institution's Calisphere gallery. 3) OAC "subject" separated into "Subject" and "Place." It should be noted that only one of the three places listed are actual geographic locations.

In the Calisphere version of "The Wild Cat,"[74] the lack of information about the source collection makes it more difficult to determine what the image is about and where it came from. The items in the subject field do indicate that the image was taken in San Francisco and relates in

[74] Calisphere, "The Wild Cat," *Chinese in California Virtual Collection: Selections from the California Historical Society*, https://calisphere.org/item/ark:/13030/hb3w1002x7/.

some way to Chinese American business or businesspeople, which is similar to “tradesmen”. However, the time period, name of the photographer, or fact that the image was taken in the context of a voyeuristic fascination with Chinatown are all completely missing. Especially without the explicit mention of an outsider gaze, this image may be understood as simply objective documentation of the everyday, rather than an intentional attempt to document and portray the exotic, the other. On OAC, the choice to title this image by the subject’s work rather than their name is understood as an obvious consequence of the lack of actual relationships or familiarity with the community being documented. Without critical self-reflection like in the series description in OAC, the title choice appears more mundane, yet another unnamed Chinese person in the archival sea of unnamed Chinese people as seen in the DPLA in figure 1.

Additionally, the two subjects that begin with a location “Chinatown (San Francisco, Calif.) -- Commerce” and “Chinatown (San Francisco, Calif.) -- Pictorial works” has been converted into three entries in the place field: “Chinatown (San Francisco, Calif.)”, “Commerce”, and “Pictorial works.” Only one of these three are actually a place, which can be confusing for a user who may not understand how the other two subjects ended up in the “place” field. By going to the OAC image page (figure 9, item 1) and then on to the OAC image detail page, in two clicks, the series and source collection can be found. However, because most of the metadata fields are the same between the OAC and Calisphere image pages, a user inexperienced with archival research may not be able to locate or understand the information in the “related item” field.

*Digital Public Library of America (DPLA)*

DPLA periodically scrapes metadata from Calisphere and other partner institutions so that all those materials can be searched and accessed from a single place. However, only items are ingested and there is no collection-level gallery page. In fact, the collection field from Calisphere is not even ingested into DPLA, since that information is not relevant for a platform designed precisely so that users do not need to be limited to a single collection when searching for materials. Without the collection field or any sort of collection-level gallery, there is no way to view all records from the same collection together, short of looking them up manually by

name and adding them to a user-created list, but there is also implicitly understood to be no reason to want to do that -- after all, collections in DPLA already exist as collections on other platforms, such as Calisphere, and exist in DPLA so that they can be searched in conjunction with millions of other items from all across the nation, all at once. Additionally, because archival collections are conventionally single-*fonds*, rather than topical, a collection field would usually contain information about creator rather than the subject of the records anyways. Aside from collection, all other metadata fields are ingested directly from Calisphere, and a button with the label “View Full Item” (figure 10, item 1) takes the user back to the Calisphere image page for that particular item. “Subject” in Calisphere is updated to “Subjects” in DPLA, and “Place” is renamed “Location” (figure 10, item 3). While a user can easily ascertain that the “The Wild Cat” was taken in San Francisco’s Chinatown based on the “place” field (figure 10, item 3), despite the absence of the series-level metadata that provides this same information in the OAC detail image view, similar to Calisphere, items in DPLA without subject or description -- which constitute nearly half of all items in *Chinese in California* at the point that they are ingested from Calisphere into DPLA -- are very difficult to find. With an even greater number of digital objects than Calisphere, even if a user was interested in the item above title one “[1]”, it would would be nearly impossible to find among millions of other items. Additionally, items without descriptive metadata are

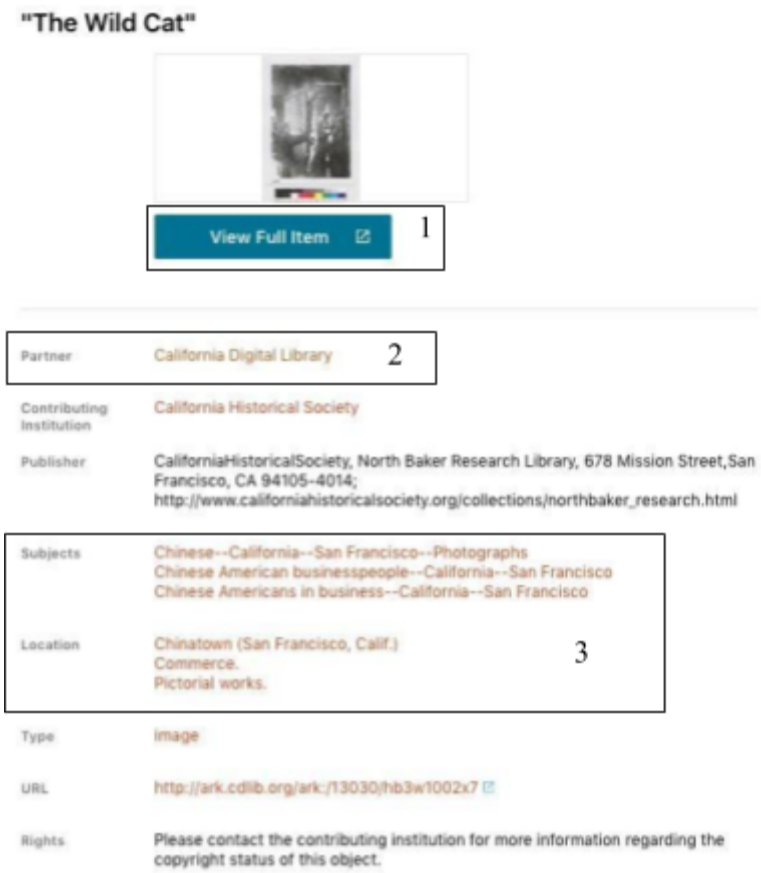



Figure 10. “The Wild Cat” on DPLA. 1) Link to Calisphere image page to view larger version of the image and additional metadata. 2) California Digital Library in the “Partner” field indicates that this item was harvested from Calisphere. 3) “Subject” and “Place” in Calisphere bcome “Subjects” and “Location” in DPLA.

very difficult for understand, even if they are found. For example, from the finding aid, Album 29 of the Roy D. Graves pictorial collection[75] consists of photographs of the Chinese community and Chinatown in San Francisco, including photograph 31, titled "A Back Yard." However, because that item in DPLA has no subject, description, source collection, or collection information listed, it would be unlikely for a user to know that it is an image of a backyard in Chinatown. In Calisphere, the user would at least see that the item is part of the *Chinese in California* collection thus likely related to Chinese Californians in some, but in DPLA no collection is listed. In other words, that same item would be understood in OAC as a backyard of a Chinese family in San Francisco's Chinatown, and in DPLA as simply a generic image of a backyard. The user can of course click through to the Calisphere image view, and then the OAC image views, but the small thumbnail and lack of information may make the user less likely to investigate further, since they are unable to determine whether or not the record is relevant to the interests.

While DPLA does have the advantage of aggregating items across platforms and allows filtering based on partner, contributing institution, subject, location, filtering by these fields all rely on exact string matches. Unless all partner institutions are using the same controlled vocabulary, clicking a subject will not actually return all items about that subject -- only those that are states in exactly the same way. Long compound subjects are not broken down into their constituent parts, so clicking on the subject "Chinese Americans in business--California--San Francisco" will only return items with that exact string. There is no way to select just "Chinese Americans in business" aside from copy-pasting it into the search bar. Consequently, clicking on the subject of items in this collection mostly return other items from the same collection, since they all share a controlled vocabulary. There is also no validation for what information can go into different fields, so the non-place places are also ingested from Calisphere into DPLA as location, making matches on location confusing for users.

[75] Roy Daniel Graves, *Roy D. Graves pictorial collection* (BANC PIC 1905.17500--ALB), the Bancroft Library, University of California, Berkeley.

| | | | | Contextual Metadata Labels | | | |
|---|---|---|---|---|---|---|---|
| | **Collection-level Description** | **Display of Items** | **Individual Item View** | **Context of Record Creation** | **Subject of Record** | **Links to other platforms** | **Degree of Aggregation** |
| **Bancroft Project Websites** | Financial, technical, and archival processes explained, only place with information about Oroville Chinese Temple | *None* | | | | All pages link to OAC finding aid | None, two separate websites designed for one project |
| **OAC** | Full heirarchical finding aid with collection- and series-level information | Gallery or within the finding aid container list, organized by series, subseries (contribuing institution), and source collection with metadata at each of those levels | Image view with large image and sparse metadata; image detail view with full item-level metadata, including source collection, series, and a link back to the finding aid | "Scope and content note", may exist at source collection- or item-level | Listed in finding aid as "Subject and Indexing Terms" and in the item detail view as "Subject", may exist at source collection- or item-level | *None* | CDL's aggregator for finding aids throughout California, as well as older archival materials |
| **Calisphere** | Abstract only | Four separate galleries for the four contributing institutions | Image view with all metadata for OAC except the series and source collection; places automatically extract from subjects and "scope and content note" changed to "description" | "Description", item-level only | "Subject"; "Place" is automatically generated from "Subject" fields that begin with geographic locations, "Place" may contain non-geographic information; item-level only | Gallery view links to OAC finding aid; Bancroft selections links back to Bancroft project-specific site; Calisphere image view links to OAC image view | CDL's aggregator for archival, bibligraphic, and other digital objects across the UC sytem and the state |
| **DPLA** | *None* | No way to view all items from the collection together; search and browse interface allows access to all DPLA materials at once | Image view with all metadata from Calisphere, except the particular collection | "Description", item-level only | "Subject"; "Location", item-level only | DPLA image view links back to Calisphere image view | US-wide aggregator for archival, bibliographic, and other digital objects |

Table 3. Overview of the collection- and item-level description, item arrangement, contextual metadata labels, inter-platform links, and degree of aggregation for each platform

*Linkages Across Platforms*

The *Chinese in California* virtual collection works to address the semantic harm caused by the absence of widely accessible materials and inaccurate and derogatory representation of Chinese Americans in the archive. This semantic harm impacts members of the Chinese American community, who suffer the negative effects of symbolic annihilation in the archival realm, as well as researcher and the broader community, who are not able to learn about or understand the community accurately. Because archival access and representation is mediated so heavily by metadata as the way that contextual information about the record is conveyed, *Chinese in California* not only seeks to make records available through digitization and online hosting but also strives to improve the archival metadata attached to the records in order to contextualize and enrich the meaning of those records. Providing detailed information about the source collection records are selected from makes explicit the provenance of the record by making known the context of record creation. Organizing the series thematically also serves to complexify notions of archival provenance by providing a better understand of the historical and cultural currents that gave rise to the types of records found within the collection. Being explicit about the context of the records is especially important for derogatory or offensive records within the collection because it acknowledges and addresses the semantic harm caused by the archives choice to collect and make available such materials.

The Bancroft's project-specific website is only platform that is explicit about the technical and archival processes undertaken in the creation of this collection. It is also the only platform with information about the Oroville Chinese Temple materials. However, this website does not have detailed information about the series or organization of the actual records within the collection, and cannot be used to access the materials themselves. Because the OAC finding aid contains the most detailed contextual information on the collection-, series-, source collection-, and item-level, it serves as the de facto guide to the collection itself, and all platforms have direct links for the users to find more information through OAC, except DPLA, where the user must take an additional step through Calisphere first. Though the OAC finding aid is the most complete single source of contextual information about the collection,

information from the explicit discussion of archival and technical processes in the Bancroft site, as well as information about the Oroville Chinese Temple materials were never added to the finding aid, and the finding OAC finding aid does not link back to the Bancroft's project site, rendering this valuable contextual information silent and invisible. Having contextual information across two separate platforms that do not interface bi-directionally with each other makes it difficult to get a complete picture of the full historical and archival context for the records.

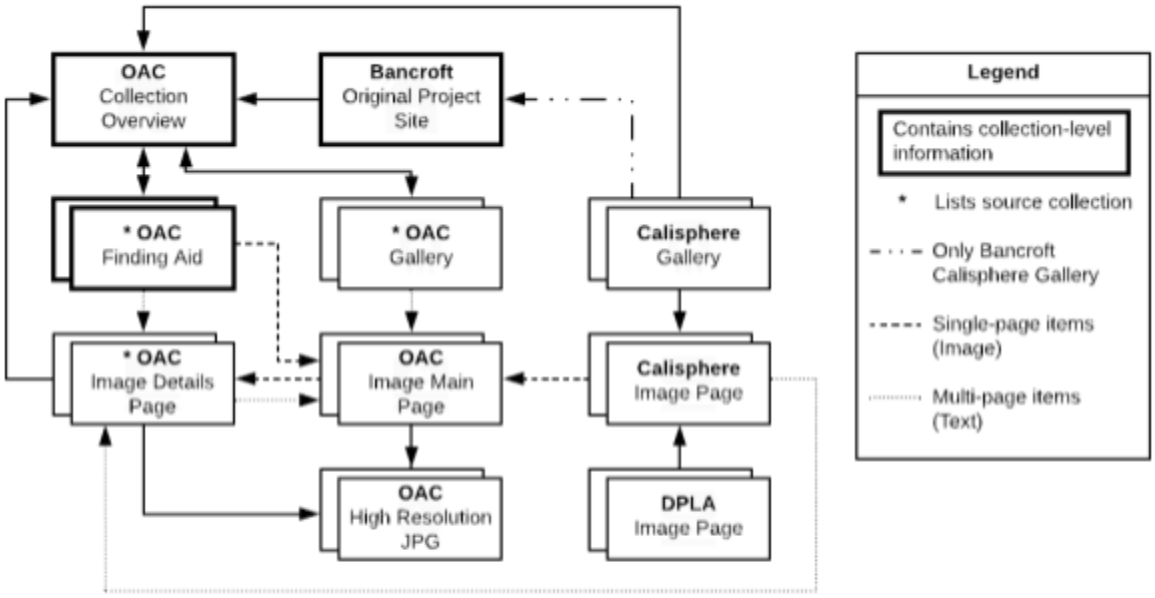



Figure 11. The linkages between the various pages of each platform. Because the most complete contextual information at the collection- and item-level exist in the OAC finding aid (though information about the technical and archival processes remain only on the Bancroft site), all other platforms ultimately point back to the finding aid. However, it is important to note that the more aggregate the platform, the more steps it takes to a user to navigate to the finding aid.

Both the Bancroft's project-specific website and OAC are platforms geared towards archival representation, and thus archival metadata at the collection-level. However, Calisphere and DPLA are both designed as item-level repositories, and can only ingest and display item-level records and metadata. However, these platforms aggregate a wider range of material types and a greater number of total items, which provides access to more materials in a single place. Additionally, in contrast to the hierarchical and rigid arrangement of records found in OAC, both Calisphere and DPLA allow for more customizable ways for users to view and

organize records according to their own needs and interests. However, because much of the contextual information created for the Bancroft project website and the OAC finding aid exist as metadata above the item-level, significant historical and cultural context for the creation of the collection and the of the records within and information that helps place individuals records within their archival bond is lost in the translation of records to Calisphere and DPLA. This loss of context is especially severe for materials from the Ethnic Studies Library, where active collecting around marginalized communities and a social justice orientation to the institution allowed them to provide the most thorough source collection level metadata -- none of which was transferred to Calisphere or DPLA, leaving those items description-less, subject-less, and difficult to find.

Though all platforms ultimately connect back to the OAC finding aid, which contains most of the contextual information, the more aggregate the platform or more item-specific the page, the further a user must travel to reach the finding aid. These platforms and pages have the lease contextual information, and thus would benefit the most from easy navigation to the finding aid, where that information can be found. It is the most difficult to access contextual information from DPLA, since the DPLA image page has the least contextual metadata, and thus the greatest need for additional information to better understand the record, but is also the most steps away from the OAC finding aid, making the finding aid harder and more work to successfully navigate to. More highly aggregated platforms allow for greater flexibility for users to customize their experience of the record access and for a wider the array of materials found on one single platform. However, conversely, these platforms have less contextual information, which can make records more difficult to find and understand accurately, and are further from the OAC finding aid, making it more work to navigate to the page where contextual information is provided.

**Limitations**

The primary limitations of this study were related to case selection. In particular, despite the long history and relatively large size of the Chinese diaspora in the United States, there was only one collection that met the case selection criterion of having archival description and metadata, and for a variety of reasons the materials in this collection may not be representative of the majority of digitized materials now available on aggregate platforms. The *Chinese in California* is also something of an unusual archival case study, since it a virtual topical collection that draws on materials from multiple *fonds*, archives, and institutions, rather than a single archival collection in the traditional single-provenance sense. However, it is extremely aware of and intentional about conventional archival processes and makes clear and intentional choices to follow or deviate from them, making it a good just not typical archival case study. Furthermore, because *Chinese in California* was created as a digital project rather than a digitization of an existing physical collection, significant additional metadata was created with digital display in mind and only some of the metadata was translated from the pre-existing physical format. This collection was also created in the very early days of online digitization, it is likely that digitization and metadata standards have changed since then.

There were also certain issues that are related to metadata and semantic harm that were outside the scope of the case study. The first was an assessment of the quality of the existing metadata. While the whole collection was assessed for the presence or absence of various contextual metadata fields at the item- and source collection-levels, there was no similarly thorough assessment for the quality of the metadata present. While metadata quality was examined for particular items in the case study, there was no formal process of identifying the prevalence of problems across the collection. For example, while it was noted in the findings that sometimes automatically generated “place” information in Calisphere included subject keywords that were not geographic places, there was no assessment of how often that or other errors occurred across the collection as a whole. Additionally, since this case study examined the presence rather than quality of metadata, an item having any information in the “place” field in Calisphere was considered sufficient to count as that record having a subject access point, despite the fact that ambiguous or misleading such as non-place subjects in the “place” field

could be confusing for users. There was also no examination of whether or not certain qualities about a record, such as language, item type, or whether it was collected by or about the community, affected the quality of metadata or the prevalence of errors. Because these factors are difficult to identify without coding each item manually, they were not explored within the context of this study.

Because of the case study methodology used, a single archival collection is examined in-depth to illustrate issues at play in the broader intersections of community archives, topical collections, and digital aggregation rather than to prove a particular hypothesis or identify the effect of a specific variable. Because of the particular case selected, this study focuses on the Chinese American community. Additional case studies on collections related to other communities marginalized on the basis of race, ethnicity, gender, sexuality, class, or other factors would help further illuminate whether archival attempts at addressing semantic harm in other communities use similar strategies or face similar challenges when integrating with aggregate platforms. Furthermore, while this collection did have four different contributing institutions, because UC Berkeley's Bancroft Library lead the project, technical and archival processes were determined by and followed the standards of a large, well-resourced institution. It is unclear what level of community control or input there was in the project, and the most community-based organization, the Oroville Chinese Temple, is the least integrated into the project. Additional study with community archives would be useful to determine how well records and collections from community archives, which exist as alternative institutions to address semantic harm done by more traditional archival institutions, integrate into digital and aggregate platforms.

Lastly, while the Bancroft website was explicit about the reasoning and goals behind the archival and technical processes for the project, I was unable to find information about the rationale or policy for collection- and item-ingest of the other platforms or about why the Bancroft chose to have the collection displayed across so many different platforms. Because of the age of *Chinese in California* -- two decades at this point, finding people familiar with the collection or even the processes at the time the collection was created proved too difficult. I did interview someone at CDL, but they noted that *Chinese in California* was arranged and

described quite differently from most newer collections and were not able to give me much information beyond that. I also attempted to access records related to the creation of the collection and its subsequent ingest into various platforms, but found the Bancroft's own records not to be catalogued or requestable. Interviews with relevant staff members and access to internal organizational records would help provide more insight into collection and platform goals and decision-making.

**Implications**

The *Chinese in California* virtual collection seeks to make visible the records of the Chinese diaspora in California, serving as a corrective measure to the semantic harm done by the historic erasure and misrepresentation of that community within archival institutions. The project not only makes a wide range of records available as digitized items viewable online by anyone with an internet connection, but also seeks to confront and address rather than minimize the semantic harm caused by historically derogatory and offensive materials. The collection- and series-level description expands notions of provenance to provide social and political context for each item, beyond the traditional narrow provenance of the donor and the *fonds*. However, tracing this collection across four platforms, each with a different approach to presenting archival materials to users, we can see how contextual information is often portrayed through collection-, series-, or source collection-level descriptive metadata and how transformations of format and metadata change how materials can be understood. Because item-based aggregate systems make very different assumptions about the causes and best corrective measures for semantic harm, the technical transformation of what metadata fields remain, are changed, or are omitted completely during the ingest of materials from an archival platform to an item-based one results in an unintentional semantic transformation that strips the items of much of their broader archival and historical context, undoes the recontextualization of those materials, and reinforces semantic harm.

An archival approach to addressing semantic harm towards a community that has been erased and misrepresented in the archives requires acknowledgement of the role archivists and archives have played, before adding additional archival and historical information to recontextualize materials and allow them to be understood in more socially responsible ways. For example, the Bancroft's original project website explains why the collection uses a topical rather than *fonds*-based arrangement of materials. Instead of relying only on a narrow understanding of provenance, centering Arnold Genthe as the photographer of Chinatown, a more expansive understanding of the cultural and political currents of the time allows that *fonds* to be placed in the context of American voyeurism towards Chinatown and active Chinese

engagement with American tourists and photographers as an additional source of income. This allows the collection to present derogatory or racist materials as a part of the historical record, while also acknowledging the both the semantic harm caused by those records and the agency of the community who are the subject of the records. Collection-wide description highlights the historical and ongoing role of the archives, series-level description provides historical context for the creation of records and for the scenes and events portrayed within the records, and source collection metadata portrays each item within its archival bond among the items in its *fonds*. In the finding aid format, hierarchical metadata places each item within expanding layers of context, and illustration of the records continuum model[76] that both preserves and enhances the original meaning the record. On a technical level, this contextual information is usually recorded above the item-level, in collection-, series-, or source collection-level metadata fields so that all items can inherit the description from higher levels and be understood in its archival and historical bonds.

While an archival approach addresses the semantic harm of misrepresentation, item-based and aggregate systems operate on the assumption that the underlying problem is lack of archival access -- items are either not available because they are restricted to a single physical location or some small corner of the internet that makes them difficult to find. The solution to lack of access is interoperable item-based records, which can be aggregated into a single platform for users to more easily access more records at once -- materials from institutions throughout California for Calisphere and throughout the United States for DPLA. Because interoperability of records is the most important trait to maximize access, metadata fields that are collection-specific are a hindrance -- items must be freed from their restrictive limited contexts so that they can be found more easily. Calisphere and DPLA drop the source collection and collection from the item-level record because that information is not usually relevant in an aggregate platform where the user is searching through all collections simultaneously for specific items regardless of soruce. While OAC attempts to aggregate collection- and item-level access, with search interfaces to look for relevant finding aids or items, though not both simultaneously, Calisphere and DPLA are geared for item-level search only. Item-level organization is meant to

[76] Sue McKemmish, "Placing Records Continuum Theory and Practice;" Frank Upward, "Modelling the Continuum as Paradigm Shift in Recordkeeping and Archiving Processes and Beyond."

be more flexible for the user, where instead of looking for a topical collection, a user can search a topic directly and find all items across all collections across that item, a individualized process of continuous classification[77] that does not rely on the archivist.

However, because aggregate platforms like DPLA only scrape item-level records from existing institutional platforms like Calisphere, which shares item-level records with a platform like OAC, which stores much of its metadata at the collection- and not item-level, the item-level records ingested in DPLA are now much harder to find. As mentioned earlier, while only a quarter of the items in OAC have no subject or scope and content note -- though these have at least some context provided by the name of the source collection, which tends to include creator and dates of creation, nearly half the items in Calisphere and DPLA have no subject or scope and content note. Without the source collection information or even the abstract for the collection, these items are difficult to search for and difficult to understand when discovered. The automatic conversion of metadata into more specific fields, for example dividing subject into subject and place or separating out creator/collector into separate two separate fields, can also make the record more ambiguous, since there are no restrictions on whether or not the metadata is appropriate for that field, leading to confusing cases like non-location words in the location field. The more aggregate the platform, the greater the number of steps to the user must take to find collection-level information, and the less information provided to help the user with that process.

Because collection-level archival and item-level aggregate platforms have different understandings of the underlying semantic harm and consequently the ideal solution, though materials can be translated and ingested from institutional to aggregate platforms, the semantic impact of this technical transformation undoes the archival corrective and reinforces semantic harm against the marginalized community who is the subject of the records. This project was undertaken because of the lack of appropriate archival representation about the Chinese diaspora in California in the late nineteenth- and early twentieth- centuries. Before the project, many records from this time period existed within archival institutions but were not easily findable, since they were often part of larger collections about other topics -- for example, a few images of Chinatown are easily lost within a large photographic collection of San Francisco. The

[77] Bak, “Continuous Classification.”

description and context created to make these materials findable and understandable in OAC, but after item-only ingest into DPLA, now half of these items once again sit mute, existent but almost impossible to actually find.

Additionally, the collection seeks to address not only the lack of representation but the historic misrepresentation and derogatory mischaracterization of the community by providing historical and archival context for those materials. However, these items stripped of context become once again simply harmful derogatory misrepresentations of the Chinese community. And furthermore, these derogatory depictions are now digitized and disseminated online beyond the confines of the physical archive, spread without warning or context. Because institutions who focus on social justice and addressing semantic harm have the interest and subject expertise to provide more contextual information beyond the item description, their work is disproportionately erased in the translation from institutional to aggregate platforms. For example, the Ethnic Studies library, which has more materials from the community itself, provided rich source-collection level subject and scope and content notes for all material it contributed. Because all of this metadata was provided on the source collection- and not item-level, none of its records in DPLA have any subject or description (the Calisphere and DPLA term for the scope and content note), rendering them invisible. The erasure of context in the translation of materials from collection-level to item-level platforms re-buries both the records and the contextual education, causing renewed semantic harm.

**Further Areas of Research**

The inability of aggregate systems to handle archival information well, disproportionately affects archival collections about marginalized communities, where historical and archival background is more important for providing contextualization for the records. However, the disconnect here is not a purely technical problem of metadata format or platform design. Rather, difficulty integrating culturally sensitive and specific materials into digital platforms often occurs because of the separation of those parts of the archival profession into separate, mutually exclusive fields. In other words, people who care about the social and political implications of archives go into community archives or archival education and not technical fields, and metadata archivists are taught about the technical specifications but not ethical implications of platforms and processes. The critical areas of archival studies have spent substantial time and energy exploring the ways that archives as a part of Western scientific and academic projects have marginalized and caused harm to certain communities, but do not dialogue with the technical areas of archival studies, which are geared towards technology as progress and as panacea. This issue is not simply a lack of diverse perspectives and practitioners in more technical archival fields, but the fact that technical specializations are framed in a way that seems irrelevant to those interested in social justice and reparative archiving.

In order to address this disconnect between the parts of archival studies that deal with addressing harm to marginalized communities and with the technical details of digital platforms, there needs to be additional work at this intersection, research that is equally informed by critical theory and technical expertise. I have identified several areas of research, which fall broadly into two categories: programmatic solutions to metadata assessment and remediation and the role of aggregate platforms.

Generally, archival institutions focus on marginalized communities as part of term-limited project, working with subject specialists to create a time- and labor-intensive collection, like the *Chinese in California*, which selects and recontextualizes thousands of items from hundreds of source collections. However, for institutions with large holdings, relying on term-limited projects and periodic community engagement is extremely slow-going and cannot

realistically cover all materials held by the institutions. Additionally, the relegation of social justice and reparative archiving to special projects excludes those areas from consideration in the central workflow and infrastructure of the archival institution. Additional research on programmatic archival assessment at scale would be useful in determining the full of extent of archival holdings, metadata quality, and areas problem areas. Because of the technical standards needed to ingest materials into different platforms, institutions already have programmatic tools that determine whether metadata is present and well-formed for required fields. However, tailoring those tools to assess the more semantic or qualitative aspects of existing metadata would help institutions identify areas where existing metadata is inadequate or inappropriate. For example, a more general version of Geraci's UC Riverside-specific tool[78] that identifies where outdated or offensive terminology for racial or ethnic groups could help identify collections and items that need to be updated, or at least given additional historical context. This is a more thorough way of assessing the collection than waiting for emails from disgruntled users who have stumbled across particular items. Lack of information is one of the biggest challenges to creating and implementing successful interventions.

Once problematic areas have been identified, programmatic methods for remediation would make addressing these issues much more feasible. Enlisting the help of machines can help make tedious tasks less time-intensive and problems can be fixed simultaneously across all holdings. For example, a thesaurus with preferred and non-preferred racial and ethnic terms can be used to update or add the preferred terms to older archival records. Topic modelling, natural language processing, or other forms of machine learning and data-mining can also be used to try and generate subject access points to improve the findability and intelligibility of records or to attempt to find information for missing metadata fields. If there are concerns about algorithms making bad updates, tools can require archivist approval before actually making a change.

While aggregate platforms like DPLA and the European Union's equivalent, Europeana, currently passively scrape and ingest materials from partner institutions, who are expected to provide and quality control the content, additional exploration into the aggregator as a site of metadata assessment or remediation would help address some of the issues around the

[78] Noah Geraci, "Programmatic approaches to bias in descriptive metadata," *Code 4 Lib* (San Jose, 20 February 2019).

heterogeneity of archival practices and materials. Assessment at the point of ingest by the aggregator would allow comparison between institutions to determine common issues and best practices. This could be especially useful to smaller institutions that may not have the technical expertise to create and run such an assessment themselves. Remediation at the aggregate level would also ensure that all ingested items were updated in a consistent way to maximize interoperability and usability. Because aggregate platform search engines rely on string matching, having standard spellings and formats for controlled vocabulary terms across all collections maximizes subject-access navigation. Remediation at the point of ingest would also allow the aggregate platform to use optimized metadata but allow institutional platforms to follow local rules, if desired, or updated metadata could be provided back to partner institutions to allow them to change their records, as well.

**Conclusion**

Both topical archival collections and aggregate digital platforms attempt to address the issue of archival access -- that users cannot find the materials that they are interested in. Topical archival collections use alternative arrangement strategies and provide rich contextual information in order to complexify provenance beyond the immediate creator of records to the agency and historical circumstances of the communities portrayed in the records. In conjunction with making items available online, adding additional archival metadata helps make these records more easily findable and understandable. Aggregate platforms, on the other hand, rely on interoperable item-level records as the great equalizer, allowing all records to be accessed from a single more level playing field. In other words, if traditional notions of provenance are limiting archival access and understanding, the community archives and thematic collections approach is to enrich and expand provenance to be more inclusive, while the aggregate platform approach is to get rid of it by setting up a system where the source of a record does not matter and all users can access all records equally. Aggregate platforms build off the work of institutional platforms and collections to scrape and index metadata and thumbnail, allowing those items to reach a wider audience while still respecting the institutions' authority over the content and description of the records. Though collections are translated technically seamlessly between platforms, from institutional to aggregate, the different underlying goals and logics of the platforms create problems of semantic translation, where significant meaning is lost between the institutional and aggregate platform because contextual information provided on the collection-level cannot follow the record into a platform that only ingests items.

Many older digital collections, such as *Chinese in California*, were created before the rise of web scraping and aggregation and were not designed with those technologies in mind. Though *Chinese in California* is now available in institutional and aggregate, older and newer platforms, it seems a though its metadata was not significantly reassessed and restructured aside from ensuring conformation to technical standards for ingest. Many archival platforms are adapted from systems built for libraries or inspired by networked open web platforms like Google, which is not designed to handle hierarchical metadata like in a finding aid, display the links between items that signify the archival bond, or to provide contextual information about the circumstance

of creation for a record. Consequently, while the aggregate platform is meant to expand the reach of materials in the archival platform, the materials are translated with much less information and context. In order for materials to be transferred successfully from archival to aggregate platform, with contextual information intact, their item-level metadata would need to be updated to inherit the information from higher levels in a more explicit way, such as making sure that all source-collection information is listed in a separate item-level scope and content note, or the platform would need to be changed to be able to represent different levels of metadata and the relations between items. For the latter to occur, aggregate platforms would need to come to see provenance and its description as integral to the meaning of the record, rather than as a barrier to access that causes unnecessary siloing and fragmentation of materials.

Enriching the item-level metadata is important not only to help users of aggregate platforms find and understand materials, but would also allow aggregate platforms to better achieve their goal of allowing flexibility and customizability for users, Weinberger's third order or order[79] where users are not limited to the organization of collections imposed by physical limitation and by archivists. This has the potential to be especially useful for multiply marginalized communities, where the grouping of materials would not have to be restricted to a single axis of identity -- materials about Angela Davis could show up under materials about Black history, women's history, and communism: a complex, intersectional social and political provenance.

[79] Weinberger, *Everything is Miscellaneous*.

**References**

Adler, Melissa. "The Case for Taxonomic Reparations." *Knowledge Organization* 43, no. 8 (2016): 630–40.

———. "Classification Along the Color Line: Excavating Racism in the Stacks." *Journal of Critical Library and Information Studies* 1, no. 1 (January 29, 2017): 1–32. https://doi.org/10.24242/jclis.v1i1.17.

———. "'Let's Not Homosexualize the Library Stacks': Liberating Gays in the Library Catalog." *Journal of the History of Sexuality* 24, no. 3 (September 2015): 478–507. http://dx.doi.org/10.7560/JHS24306.

Bak, Greg. "Continuous Classification: Capturing Dynamic Relationships among Information Resources." *Archival Science* 12, no. 3 (2012): 287–318. http://dx.doi.org.proxy.lib.umich.edu/10.1007/s10502-012-9171-8.

The Bancroft Library, "About the Project," Chinese in California, 1850-1925. December 3, 2008. Accessed March 20, 2019. http://bancroft.berkeley.edu/collections/chineseinca/about.html.

———. "Home." Chinese in California, 1850-1925. December 3, 2008. Accessed March 20, 2019. http://bancroft.berkeley.edu/collections/chineseinca/index.html.

———. "Oroville Chinese Temple." December 11, 2006. Accessed March 20, 2019. http://bancroft.berkeley.edu/collections/oroville/.

Bearman, David A, and Richard H Lytle. "The Power of the Principle of Provenance." *Archivaria*, no. 21 (Winter 1985-86): 14–27.

Bowker, Geoffrey C., and Susan Leigh Star. *Sorting Things out: Classification and Its Consequences*. MIT Press, 1999.

Brothman, Brien. "Orders of Value: Probing the Theoretical Terms of Archival Practice." *Archivaria*, no. 32 (Summer 1991): 78–100.

Calisphere. "Chinese in California Virtual Collection: Selections from the Bancroft Library." Accessed March 20, 2019. https://calisphere.org/collections/143/.

———, "The Wild Cat." Chinese in California Virtual Collection: Selections from the California Historical Society. Accessed March 20, 2019. https://calisphere.org/item/ark:/13030/hb3w1002x7/.

Caswell, Michelle. "'The Archive' Is Not an Archives: On Acknowledging the Intellectual Contributions of Archival Studies". *Reconstruction* 16, no. 1 (2016), unpaginated.

———."Using Classification to Convict the Khmer Rouge." *Journal of Documentation* 68, no. 2 (2012): 162–184. http://dx.doi.org/10.1108/00220411211209177.

Caswell, Michelle, Marika Cifor, and Mario H. Ramirez. "'To Suddenly Discover Yourself Existing': Uncovering the Impact of Community Archives." *The American Archivist* 79, no. 1 (June 1, 2016): 56–81. https://doi.org/10.17723/0360-9081.79.1.56.

Christen, Kimberly. "Archival Challenges and Digital Solutions in Aboriginal Australia." *SAA Archaeological Record* 8, no. 2 (August 2008): 21-24.

———. "Opening Archives: Respectful Repatriation." *The American Archivist* 74, no. 1 (Spring/Summer 2011): 185-210. https://doi.org/10.17723/aarc.74.1.4233nv6nv6428521.

Dalbello, Marija. "A Phenomenological Study of an Emergent National Digital Library, Part I: Theory and Methodological Framework." *The Library Quarterly* 75, no. 4 (October 1, 2005): 391–420. https://doi.org/10.1086/502784.

Davey, Lynn. "The Application of Case Study Evaluations." *Practical Assessment, Research & Evaluation* 2, no. 9 (November 1991): 1-2. https://pareonline.net/getvn.asp?v=2%26n=9

Derrida, Jacques. *Archive Fever: A Freudian Impression*. University of Chicago Press, 1996.

Douglas, Jennifer. "Origins: Evolving Ideas about the Principle of Provenance." In *Currents of Archival Thinking*, edited by Terry Eastwood and Heather MacNeil. Santa Barbara, CA: Libraries Unlimited, 2010: 25-52.

The DPLA Archival Description Working Group. "Aggregating and Representing Collections in the Digital Public Library of America." November 2016. Accessed March 20, 2019. http://bit.ly/dplaCollections.

Fidel, Raya. "The Case Study Method: A Case Study." *Library and Information Science Research*, no. 6 (1984): 274-288.

Foucault, Michel. *The Archaeology of Knowledge*. Translated by A. M. Sheridan Smith. London and New York: Routledge, 2002.

Furner, Jonathan. "Conceptual Analysis: A Method for Understanding Information as Evidence, and Evidence as Information." *Archival Science* 4 (2004): 233-265. https://doi.org/10.1007/s10502-005-2594-8.

Genthe, Arnold. "The Wild Cat." Photos: San Francisco--Chinatown (1895-1906): Arnold Genthe -- Photos No. 6 (Tradesmen): FN-02252. California Historical Society, San Francisco.

Geraci, Noah. "Programmatic approaches to bias in descriptive metadata." Code 4 Lib. San Jose, 20 February 2019.

Gilliland-Swetland, Luke. "The Provenance of a Profession: The Permanence of the Public Archives and Historical Manuscripts Traditions in American Archival History." *The American Archivist* 54, no. 2 (Spring 1991): 160-175.

Graves, Roy Daniel. Roy D. Graves pictorial collection: BANC PIC 1905.17500--ALB. The Bancroft Library, University of California, Berkeley.

Hensen, Steven L. "The First Shall Be First: APPM and Its Impacts on American Archival Description." *Archivaria* 35 (Spring 1993): 64–70.

Jenkinson, Sir Hilary. *A Manual of Archive Administration*. London: Percy Lund, Humphries & Co., 1937.

Kelleher, Christian. "Archives Without Archives: (Re)Locating and (Re)Defining the Archive Through Post-Custodial Praxis." *Journal of Critical Library and Information Studies* 1, no. 2 (July 7, 2017). https://doi.org/10.24242/jclis.v1i2.29.

Ketelaar, Eric. "Tacit Narratives: The Meaning of Archives." *Archival Science* 1, no. 2 (2001): 131–141. https://doi.org/10.1007/BF02435644.

Lougheed, Brett, Ry Moran, and Camille Callison. "Reconciliation through Description: Using Metadata to Realize the Vision of the National Research Centre for Truth and Reconciliation." *Cataloging & Classification Quarterly* 53, no. 5–6 (July 4, 2015): 596–614. https://doi.org/10.1080/01639374.2015.1008718.

Luker, Trish. "Decolonising Archives: Indigenous Challenges to Record Keeping in 'Reconciling' Settler Colonial States." *Australian Feminist Studies* 32, no. 91–92 (April 3, 2017): 108–25. https://doi.org/10.1080/08164649.2017.1357011.

Lynch, Shola. "From Filmmaker to Curator: The Politics of Archiving." The Humanities and the Arts in the Integrated Knowledge University: Opening the Archive. Philadelphia: University of Pennsylvania, 13 September 2014.

McKemmish, Sue. "Placing Records Continuum Theory and Practice." *Archival Science* 1 (2001): 333-359.

Mckemmish, Sue, Shannon Faulkhead, and Lynette Russell. "Distrust in the Archive: Reconciling Records." *Archival Science*, no. 11 (2011): 211–239. DOI 10.1007/s10502-011-9153-2.

Millar, Laura. "The Death of the Fonds and the Resurrection of Provenance: Archival Context in Space and Time." *Archivaria*, no. 53 (2002): 1-15.

Muller, Samuel, Johan Adriaan Feith, and R. Fruin. *Manual for the Arrangement and Description of Archives: Drawn up by the Netherlands Association of Archivists.* Translated by Arthur H. Leavitt. New York: H. W. Wilson, 1968.

Nesmith, Tom. "Still Fuzzy But More Accurate: Some Thoughts on the 'Ghosts' of Archival Theory." *Archivaria*, no. 47 (1999): 136-150.

Online Archive of California. "Collection Details." Guide to the Chinese in California Virtual Collection. Accessed March 20, 2019. https://oac.cdlib.org/findaid/ark:/13030/kt5p3019m2/admin/.

———. "Collection Overview." Guide to the Chinese in California Virtual Collection. Accessed March 20, 2019. https://oac.cdlib.org/findaid/ark:/13030/kt5p3019m2/.

———. "Container List." Guide to the Chinese in California Virtual Collection. Accessed March 20, 2019. https://oac.cdlib.org/findaid/ark:/13030/kt5p3019m2/dsc/#dsc-1.8.5.

———. "Online items." Guide to the Chinese in California Virtual Collection. Accessed March 20, 2019. https://oac.cdlib.org/view?docId=kt5p3019m2;developer=local;style=oac4;doc.view=items.

———. "The Wild Cat." Chinese in California. Accessed March 20, 2019. https://oac.cdlib.org/ark:/13030/hb3w1002x7/?brand=oac4.

Riley, Jenn, and Kelcy Shepherd. "A Brave New World: Archivists and Shareable Descriptive Metadata." *The American Archivist* 72, no. 1 (2009): 91–112.

Sandy, Heather Moulaison, and Jenny Bossaller. "Providing Cognitively Just Subject Access to Indigenous Knowledge through Knowledge Organization Systems." *Cataloging & Classification Quarterly* 55, no. 3 (April 3, 2017): 129–152. https://doi.org/10.1080/01639374.2017.1281858.

Society of American Archivists. "Statement of Principles." Describing Archives: A Content Standard, Second Edition (DACS). Accessed March 20, 2019. https://www2.archivists.org/standards/DACS/statement_of_principles.

———. “Provenance.” A Glossary of Archival and Records Terminology. Accessed March 20, 2019. https://www2.archivists.org/glossary/terms/p/provenance.

Taber, Isaiah West. Souvenir of California: photographic views San Francisco, Cal. : Taber Photo., [ca.1882]: pfF869.S3.9.S718. The Bancroft Library, University of California, Berkeley.

Upward, Frank. “Modelling the Continuum as Paradigm Shift in Recordkeeping and Archiving Processes and Beyond.” *Records Management Journal* (December 2000): unpaginated.

Visvanathan, Shiv. *A Carnival for Science: Essays on Science, Technology and Development*. London: Oxford University Press, 1997.

Weinberger, David. *Everything is Miscellaneous: The Power of the New Digital Disorder.* New York: Henry Holt, 2007.

Wood, Stacy, Kathy Carbone, Marika Cifor, Anne Gilliland, and Ricardo Punzalan. “Mobilizing Records: Re-Framing Archival Description to Support Human Rights.” *Archival Science* 14, no. 3 (October 1, 2014): 397–419. https://doi.org/10.1007/s10502-014-9233-1.

Yaco, Sonia, Ann Jimerson, Laura Caldwell Anderson, and Chanda Temple. “A Web-Based Community-Building Archives Project: A Case Study of Kids in Birmingham 1963.” *Archival Science* 15, no. 4 (December 1, 2015): 399–427. https://doi.org/10.1007/s10502-015-9246-4.

Yeo, Geoffrey “Concepts of Record (1): Evidence, Information, and Persistent Representation.” *The American Archivist* 70, no. 2 (2007): 315-343. https://doi.org/10.17723/aarc.70.2.u327764v1036756q.