

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

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

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

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

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

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

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

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

### **Current State of OCR Technology**

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

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

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

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

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

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

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

### **Current State of Discovery in Digital Collections**

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

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

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

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

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

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

### **Proposed Solution**

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

#### Transparency

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

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

#### Metadata

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

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



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

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

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

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

## Generous Design

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

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

## Conclusion

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

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

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

### Works Cited

- Baker. "A Brief History of Search Engine Optimization." *Search Engine Journal*, 26 Dec. 2017, <https://www.searchenginejournal.com/seo-101/seo-history/>.
- Bates, Marcia J. *Neolithic Information Seeking: Designing Information Systems for Our Inner Hunter-Gatherer*. Information Architecture Summit 2018, Chicago, IL.
- "Browse Tags." *California Digital Newspaper Collection*, <https://cdnc.ucr.edu/?a=cl&cl=Tags.X&e=-----en--20--1--txt-txIN-----1>. Accessed 23 Feb. 2020.
- Chassanoff, Alexandra. "Historians and the Use of Primary Source Materials in the Digital Age." *The American Archivist*, vol. 76, no. 2, Sept. 2013, pp. 458–80, doi:[10.17723/aarc.76.2.lh76217m2m376n28](https://doi.org/10.17723/aarc.76.2.lh76217m2m376n28).
- Cordes, Sean. "Student Perceptions of Search Tool Usability." *Internet Reference Services Quarterly*, vol. 19, no. 1, Jan. 2014, pp. 3–32. *Taylor and Francis+NEJM*, doi:[10.1080/10875301.2014.894955](https://doi.org/10.1080/10875301.2014.894955).
- "Digital Collections." *Library of Congress*, <https://www.loc.gov/collections/>. Accessed 23 Feb. 2020.
- Drabinski, E. (2013). Queering the Catalog: Queer Theory and the Politics of Correction. *Library Quarterly: Information, Community, Policy*, 83(2), 94–111.
- Duarte, M. E., & Belarde-Lewis, M. (2015). Imagining: Creating Spaces for Indigenous Ontologies. *Cataloging & Classification Quarterly*, (53), 677–702.
- Fear, Kathleen. "User Understanding of Metadata in Digital Image Collections: Or, What Exactly Do You Mean by 'Coverage'?" *The American Archivist*, vol. 73, no. 1, 2010, pp. 26–60.

- Fedoroff, Lara, and Chris Chandler. *How Search Really Works*. <http://ux-radio.com/2019/04/search-really-works-guest-dr-marcia-bates/>. Accessed 20 Dec. 2019.
- Furner, Jonathan. “Dewey Deracialized: A Critical Race-Theoretic Perspective.” *Knowledge Organization*, vol. 34, Jan. 2007, pp. 144–68.
- G. Chiron, et al. *Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information*. 2017 ACM/IEEE Joint Conference on Digital Libraries, Toronto, ON. 2017, pp. 1–4, doi:[10.1109/JCDL.2017.7991582](https://doi.org/10.1109/JCDL.2017.7991582).
- Halarewich, Danny. “Reducing Cognitive Overload For A Better User Experience.” *Smashing Magazine*, 16 Sept. 2016. <https://www.smashingmagazine.com/2016/09/reducing-cognitive-overload-for-a-better-user-experience/>.
- Mager, Manuel, et al. *Challenges of Language Technologies for the Indigenous Languages of the Americas*. 27th International Conference on Computational Linguistics (COLING 2018), Santa Fe, New Mexico. 2018.
- “New Search Feature: Optical Character Recognition (OCR).” *NARAtions*, 9 Sept. 2019, <https://narations.blogs.archives.gov/2019/09/09/new-search-feature-optical-character-recognition-ocr/>.
- “Search.” *Europeana*, <https://www.europeana.eu/en/search?page=1&view=grid&query=>. Accessed 23 Feb. 2020.
- “Search Results.” *Calisphere*, <https://calisphere.org/search/?q=>. Accessed 23 Feb. 2020.
- Smith, David A., and Ryan Cordell. *A Research Agenda for Historical and Multilingual Optical Character Recognition*. Northeastern University, 2018.
- Srihari, Sargur N., et al., editors. “Optical Character Recognition (OCR).” *Encyclopedia of Computer Science*, John Wiley and Sons Ltd, 2003, pp. 1326–1333.

Stack, John. "Exploring Museum Collections Online: Some Background Reading." *Medium*, 6

Aug. 2018, <https://lab.sciencemuseum.org.uk/exploring-museum-collections-online-some-background-reading-da5a332fa2f8>.

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

Wang, Hsiang-An, and Pin-Ting Liu. *Towards a Higher Accuracy of Optical Character*

*Recognition of Chinese Rare Books in Making Use of Text Model*. 3rd International

Conference on Digital Access to Textual Cultural Heritage, Brussels, Belgium. 2018, pp.

15–18.

Whitelaw, Mitchell. "Generous Interfaces for Digital Cultural Collections." *Digital Humanities*

*Quarterly*, vol. 9, no. 1, May 2015.

## Bibliography

- Baker. "A Brief History of Search Engine Optimization." *Search Engine Journal*, 26 Dec. 2017, <https://www.searchenginejournal.com/seo-101/seo-history/>.
- Bates, Marcia J. *Neolithic Information Seeking: Designing Information Systems for Our Inner Hunter-Gatherer*. Information Architecture Summit 2018, Chicago, IL.
- . "What Is Browsing—Really? A Model Drawing from Behavioural Science Research." *Information Research*, vol. 12, no. 4, Oct. 2007, <http://informationr.net/ir/12-4/paper330.html>.
- Bell, Steven J. "Submit or Resist: Librarianship in the Age of Google." *American Libraries*, vol. 36, no. 9, 2005, pp. 68–71. JSTOR.
- "Browse Tags." *California Digital Newspaper Collection*, <https://cdnc.ucr.edu/?a=cl&cl=Tags.X&e=-----en--20--1--txt-txIN-----1>. Accessed 23 Feb. 2020.
- Chassanoff, Alexandra. "Historians and the Use of Primary Source Materials in the Digital Age." *The American Archivist*, vol. 76, no. 2, Sept. 2013, pp. 458–80, doi:[10.17723/aarc.76.2.lh76217m2m376n28](https://doi.org/10.17723/aarc.76.2.lh76217m2m376n28).
- Cordes, Sean. "Student Perceptions of Search Tool Usability." *Internet Reference Services Quarterly*, vol. 19, no. 1, Jan. 2014, pp. 3–32. *Taylor and Francis+NEJM*, doi:[10.1080/10875301.2014.894955](https://doi.org/10.1080/10875301.2014.894955).
- "Digital Collections." *Library of Congress*, <https://www.loc.gov/collections/>. Accessed 23 Feb. 2020.
- Drabinski, E. (2013). Queering the Catalog: Queer Theory and the Politics of Correction. *Library Quarterly: Information, Community, Policy*, 83(2), 94–111.

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

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

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

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

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

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

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

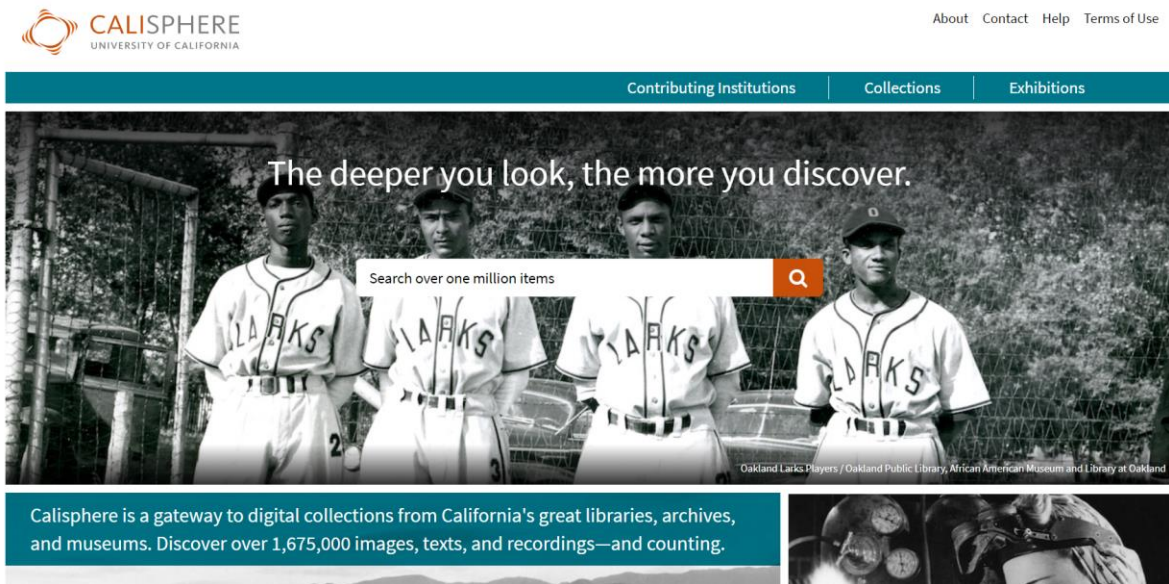
“More Guidance on Building High-Quality Sites.” *Official Google Webmaster Central Blog*, 6 May 2011, <https://webmasters.googleblog.com/2011/05/more-guidance-on-building-high-quality.html>.



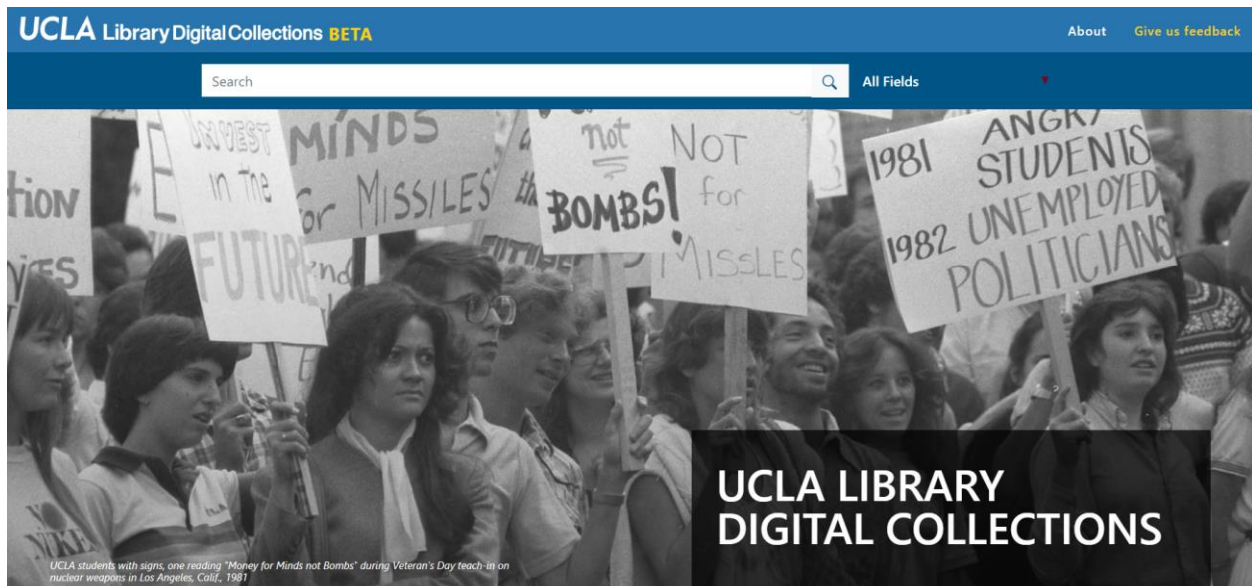
- “New Search Feature: Optical Character Recognition (OCR).” *NARAtions*, 9 Sept. 2019, <https://narations.blogs.archives.gov/2019/09/09/new-search-feature-optical-character-recognition-ocr/>.
- “Search.” *Europeana*, <https://www.europeana.eu/en/search?page=1&view=grid&query=>. Accessed 23 Feb. 2020.
- “Search Results.” *Calisphere*, <https://calisphere.org/search/?q=>. Accessed 23 Feb. 2020.
- Smith, David A., and Ryan Cordell. *A Research Agenda for Historical and Multilingual Optical Character Recognition*. Northeastern University, 2018.
- Srihari, Sargur N., et al., editors. “Optical Character Recognition (OCR).” *Encyclopedia of Computer Science*, John Wiley and Sons Ltd, 2003, pp. 1326–1333.
- Stack, John. “Exploring Museum Collections Online: Some Background Reading.” *Medium*, 6 Aug. 2018, <https://lab.sciencemuseum.org.uk/exploring-museum-collections-online-some-background-reading-da5a332fa2f8>.
- Toms, Elaine G. “Understanding and Facilitating the Browsing of Electronic Text.” *International Journal of Human-Computer Studies*, vol. 52, no. 3, Mar. 2000, pp. 423–52, doi:[10.1006/ijhc.1999.0345](https://doi.org/10.1006/ijhc.1999.0345).
- UCLA Library Digital Collections*. <https://digital.library.ucla.edu/>. Accessed 23 Feb. 2020.
- Wang, Hsiang-An, and Pin-Ting Liu. *Towards a Higher Accuracy of Optical Character Recognition of Chinese Rare Books in Making Use of Text Model*. 3rd International Conference on Digital Access to Textual Cultural Heritage, Brussels, Belgium. 2018, pp. 15–18.
- Whitelaw, Mitchell. “Generous Interfaces for Digital Cultural Collections.” *Digital Humanities Quarterly*, vol. 9, no. 1, May 2015.

## Appendix A: Digital Collections Interfaces

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



Calisphere (<https://calisphere.org/>)



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

[Add a search term](#)  [Browse](#)

Explore 58,182,351 artworks, artefacts, books, films and music from European museums, galleries, libraries and archives

[TRY THE DEMO OF THE NEW EUROPEANA](#)

Battle between Carnival and Lent, Hieronymus Bosch  
 Rijksmuseum  
 Public Domain

EXHIBITION EXHIBITION

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

**UCR** Center for Bibliographical Studies and Research

**CDNC** California Digital Newspaper Collection

Search Browse Help About

**FEATURED**

Colusa Herald 7 August 1928

**SEARCH**

**INTRODUCING A NEW LOOK AND NEW FEATURES**  
 See the Help section for more details.

**ABOUT**

This collection contains 492,311 issues comprising 5,430,664 pages and 37,145,576 articles.

The California Digital Newspaper Collection is a project of the Center for Bibliographical Studies and Research (CBSR) at the University of California, Riverside.

The CDNC is supported in part by the U.S. Institute of Museum and Library Services under the provisions of the Library Services and Technology Act, administered in California by the State Librarian.

The CBSR has received three grants from the National Endowment for the Humanities to digitize California newspapers for the National Digital Newspaper

**BROWSE**

Browse by title
 Browse by date

Browse by tag
 Browse by county

**DONATE**

Though access to the CDNC is free, maintaining and improving it is not. Please consider supporting the CDNC.

**TOP TEXT CORRECTORS**

1. Wes Keat	2,122,416
2. annh	1,252,322

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

Digital Collections

Featured Content

Historic American Buildings Survey/Historic American Engineering...

Chronicing America: Historic American Newspapers

Farm Security Administration/Office of War Information ...

Cities and Towns

Civil War Maps

Results: 1-40 of 405 | Refined by: Part of: Digital Collections Available Online

Refine your results

- Subject**
- American History 140
  - Government, Law & Politics 111
  - Performing Arts 92
  - World Cultures & History 89
  - War & Military 75
  - Local History & Folklife 56
  - Art & Architecture 45
  - Social & Business History 35
  - Photographic Prints 30
  - Portrait Photographs 30
  - More Subjects >
- Part of**
- Digital Collections x
  - Manuscript Division 94
  - Prints and Photographs Division 76

Digital Collections

View Gallery Go Sort By Select Go

**COLLECTION**  
**10th-16th Century Liturgical Chants**  
Collection Items: View 55 items

**COLLECTION**  
**Aaron Copland Collection**  
The Aaron Copland collection consists of published and unpublished music by Copland and other composers, correspondence, writings, hierarchical material

**COLLECTION**  
**Abdul Hamid II Collection**  
These photographic albums portray the Ottoman Empire during the reign of one of its last sultans, Abdul-Hamid II. They highlight the modernization of numerous aspects of the

**COLLECTION**  
**Abdul-Hamid II Collection of Books and Serials Gifted to the Library of Congress**  
Collection Items: View 323 items

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



## Appendix B: OCR Badge Wireframes

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

### *Collection-level wireframe with OCR badge*



#### About this Collection

An internal serial publication of the German News Agency (Deutsches Nachrichtenbüro) before and during the Second World War. Published three or more times a day, UCLA holdings cover the period May 8 1936 to May 25, 1940. Digitization is under way and the digital copies will be published as completed.

#### Collection Overview

Press releases, issued in newspaper form, of information released by the Deutsches Nachrichtenbüro from the early 1930s through the 1940s. Includes occasional supplements, with varying titles; for example: Deutsches Nachrichtenbüro, Deutscher Handelsdienst (1937); and Deutsches Nachrichtenbüro. Sonderausgabe (1935/1936-1938/1939).

#### Keyword Search Confidence 81%

Text within items in this collection was identified and made searchable with optical character recognition (OCR) technology. OCR technology is not always able to successfully identify text, which means some items will not successfully be retrieved with a keyword search.

Confidence ratings, or predictions for the OCR's accuracy, is noted within each item's description.

Overall, the average confidence rating for this collection is 87%.

#### Find this Collection

**REPOSITORY** University of California, Los Angeles. Library Special Collections  
**ARK** ark:/21198/zz00294nw7

#### Contact

UCLA Charles E. Young Research Library Department of Special Collections, A1713 Young Research Library, Box 951575, Los Angeles, CA 90095-1575. E-mail: spec-coll@library.ucla.edu. Phone: (310)825-4988

[Browse items in this collection](#)

### Keyword Search Confidence 81%

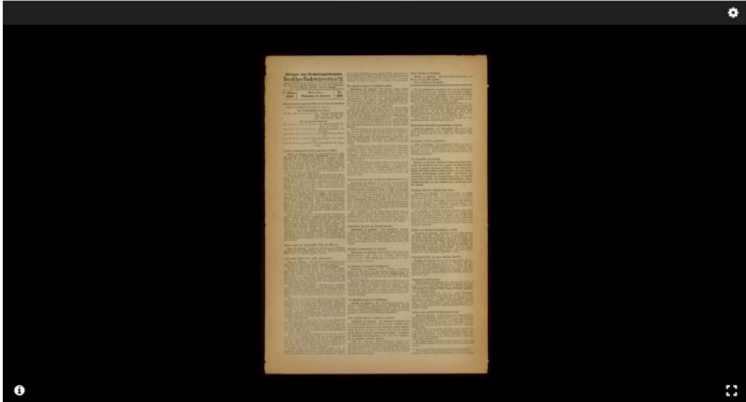
Text within items in this collection was identified and made searchable with optical character recognition (OCR) technology. OCR technology is not always able to successfully identify text, which means some items will not successfully be retrieved with a keyword search.

Confidence ratings, or predictions for the OCR's accuracy, is noted within each item's description.

Overall, the average confidence rating for this collection is 81%.

*Item-level wireframe with OCR badge*

Deutsches Nachrichtenbüro. 7 Jahrg., Nr. 153, 1940 February 14, Mittags- und Nachmittags-Ausgabe



**Item Overview**

TITLE	Deutsches Nachrichtenbüro. 7 Jahrg., Nr. 153, 1940 February 14, Mittags- und Nachmittags-Ausgabe
ALTERNATIVE TITLE	Deutsches Nachrichtenbüro Mittags- und Nachmittags-Ausgabe
LANGUAGE	German
COLLECTION	Deutsches Nachrichtenbüro

---

**Physical Description**

EXTENT	1 p.
--------	------

---

**Keywords**

GENRE	newspapers
RESOURCE TYPE	text

**Keyword Search Confidence**

72%

Text in this item was identified and made searchable with optical character recognition (OCR) technology. This confidence rating is a prediction of the OCR's accuracy. Some text may not have been identified correctly and thus will not be retrieved in a keyword search.

---

**Find This Item**

REPOSITORY	University of California, Los Angeles, Library Special Collections
LOCAL IDENTIFIER	1940-02-14_0153
ARK	ark:/21198/zz002/bw0wb

---

**Access Condition**

COPYRIGHT STATUS	unknown
LICENSE	No license recorded

## Keyword Search Confidence

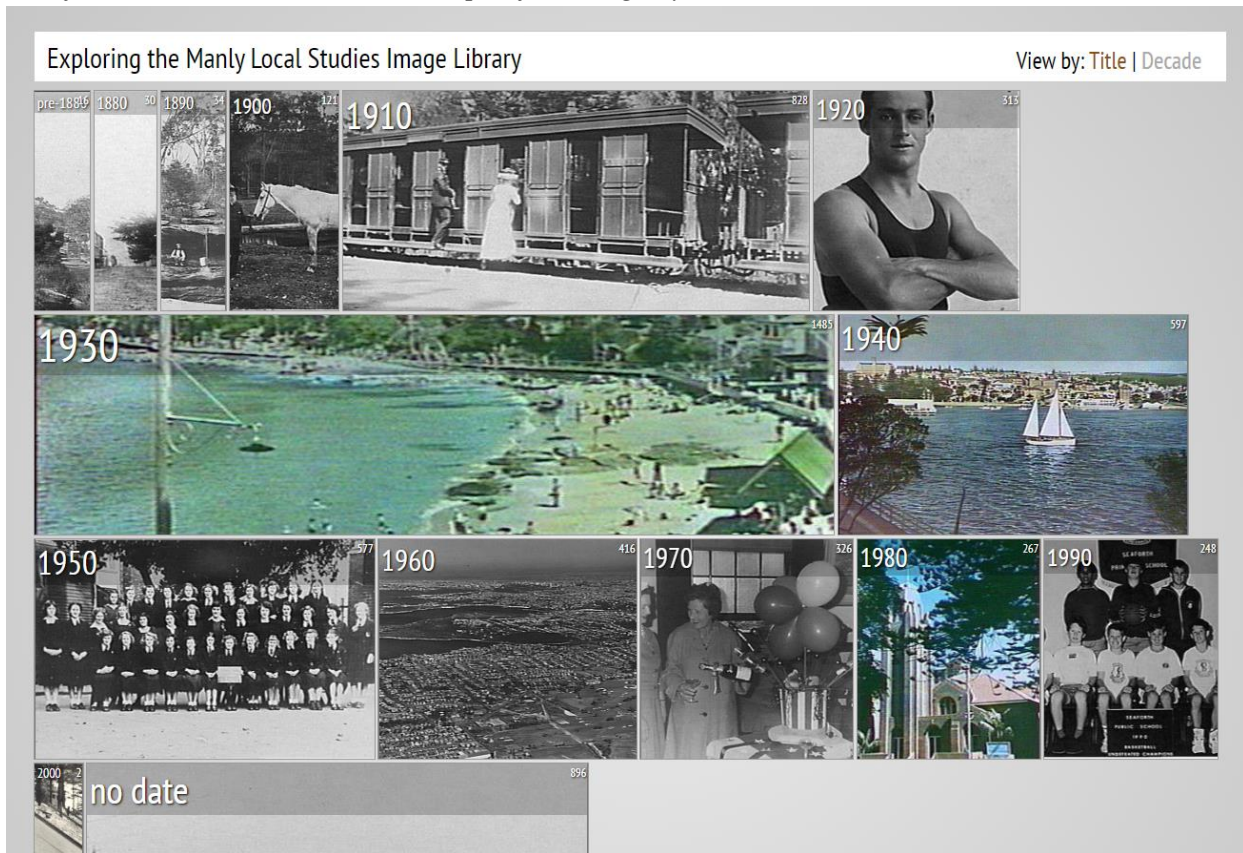
72%

Text in this item was identified and made searchable with optical character recognition (OCR) technology. This confidence rating is a prediction of the OCR's accuracy. Some text may not have been identified correctly and thus will not be retrieved in a keyword search.

### Appendix C: Generous Interface Designs

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

*Interface that communicates the scope of holdings by decade*



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

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

## STATE LIBRARY OF QUEENSLAND

The screenshot displays the 'Queenslander Mosaic Grid' interface. At the top, a navigation bar includes 'RESEARCH & COLLECTIONS', 'PLAN MY VISIT', 'DISCOVER', 'WHAT'S ON', 'GET INVOLVED', 'HOW DO I?', and 'ABOUT US'. Below this, the 'Queenslander Mosaic Grid' header is visible. A horizontal timeline at the top shows years from 1880 to 1905. A list of subjects is provided, including: ACTORS, ADVERTISEMENTS, AIRPLANES, AGED PEOPLE, AGRICULTURAL SHOWS, AVIATION, AVIATORS, BEACHES, BIRDS, BOATS, BOYS, BRISBANE, CARICATURES, CATTLE, CHILDREN, CHILDREN'S CLOTHING, CHRISTMAS, COUNTRY SCENES, CRICKET, CRICKETERS, DOGS, FARMING, FISHING, FLOWERS, GIRLS, HAIRSTYLES, HATS, HORSERIDERS, HORSES, HOTELS & TAVERNS, INDIGENOUS AUSTRALIANS, LANDSCAPES (VIEWS), MEN'S CLOTHING AND ACCESSORIES, MILITARY (PATTERNS (CLOTHES)), PIPES (SMOKING), PLANTS, PORTRAITS, RIVERS, ROYALTY, SAILING, SAILING BOATS, SHIPS, SPORT, STOCKMEN, SWIMSUITS, TREES, WOMEN, WOMEN'S CLOTHING & ACCESSORIES, WORKERS. Below the subjects is a list of names: AGNEW, GARNET, 1886-1961; ALBAN, TOM; BARKER, CAROLINE, 1894-1988; BENNETT, R. W.; BERRY; BRESSOW, LANCE; BUSTARD, WILLIAM, 1894-1973; CAMPBELL, FRANK; DALGANNO, ROY FREDERICK, LESLIE, 1910-2001; DRIVER, ADA; OSTER, J. M.; HARRIS, DOREEN; HARRISON, HARRY; HOSGAY, P. STANHOPE; LAHEY, VIDA, 1892-1988; MCBAIN, IAN; MEARS, ERNEST HAROLD, 1896-1977; MOWDEN, WILFRED; PATTERSON, BETTY; PATTERSON, ESTHER, 1892-1971; PRYNE, FRANK; JENKINS, C. E.; HED, CHARLES; RHYE, JOAN; SNEYD, WILLIAM; FORNANCE, W.; WARD, JOHN E.; WATSON, E. S.; WHITE, A. A.; WENKE, JAMES, 1908-1981. A color bar is located below the names. A central search bar contains the number '989'. The main content area features a grid of image thumbnails with dates: 16 December 1899, 30 March 1900, 15 December 1900, 11 May 1901, and a larger thumbnail for 'THE FEDERAL FLAG'.

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



Interface with mosaic tiles to facilitate scanning and browsing

The screenshot displays the Rijksstudio website interface. At the top, a banner features the text "RIJKS STUDIO" in large white letters over a dark background with a painting of a windmill. Below the banner, a navigation bar contains the text "Discover the possibilities of the masterpieces" and a button "Create your own Rijksstudio". The main content area is a grid of art tiles. The first row features three main sections: "Rembrandt van Rijn" with a large tile of a portrait and smaller tiles of other works; "Johannes Vermeer" with a large tile of a woman pouring milk and smaller tiles of other works; and "Paintings" with a large tile of a woman in a black dress and smaller tiles of other works. Below each section is a "More highlights" or "More artists" button. The second row features three sections: "Rococo" with a large tile of a portrait and smaller tiles of other works; a large central tile of a wooden sculpture of a couple; and "sketches" with a large tile of a sketch and smaller tiles of other sketches. Below each section is a "More works of art" button. The bottom of the page features a link to the Rijksmuseum website.

Discover the possibilities of the masterpieces [Create your own Rijksstudio](#)

**Rembrandt van Rijn**  
RIJKS MUSEUM  
[More highlights](#)

**Johannes Vermeer**  
RIJKS MUSEUM  
[More artists](#)

**Paintings**  
RIJKS MUSEUM  
[More works of art](#)

**Now in Rijksstudio**  
Browse 674,757 works of art and 515,310 Rijksstudios

**Rococo**  
RIJKS MUSEUM

**sketches**  
Veronika Suchodolska  
4 minutes ago - 45 works

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