49/3 2024 alj

Collection Data: Sharing, Discovery, Inspiration, and Innovation

Angela Yon

In recent years, the traditional use of digital collections as surrogates for the physical has shifted to a paradigm of viewing collections as data suitable for computational use and novel research methods. The burgeoning collections as data movement is gaining momentum among galleries, libraries, archives, and museums (GLAM) worldwide. Strategic initiatives, experimentation, innovation, and inspirational learning are occurring as digital libraries and digital humanities progress and work to develop sustainable approaches for collections as data programs. What is the position of collections as data in an ever-changing information landscape of open access, linked data, and shared data of cultural heritage collections? What has the past decade brought to the field?

Introduction

The collections as data movement has gained substantial momentum with the increase of digital collections and born-digital objects. This approach looks beyond the traditional use of digital collections as surrogates for the physical and turning digital collections and their inherent data into datasets amenable to computational use and digital humanities research methods. According to Thomas Padilla, "A collections as data paradigm seeks to foster an expanded set of research, pedagogical, and artistic potential predicated on the computational use of cultural heritage collections."[1] Increased access to datasets about cultural heritage objects and works of art creates wider accessibility, discovery, reusability, and drives knowledge, scholarship, and innovation, inspiring new perspectives. Digital collections can be transformed into different high-quality datasets and are becoming a comprehensive information space of complex data analysis and interactive data visualizations.[2] The following survey of such projects and initiatives within the past decade reveals a collections as data community evolving engagement, experimentation, and growth in a landscape of open access, linked data, and innovative data sharing.

Collections as Data: Sharing Data

Collections as data is the concept that digital collections are data that researchers can analyse using computers. It is grounded upon two scholarly trends happening concurrently: librarians and archivists have digitized an increasingly large portion of cultural heritage collections providing access to unique materials; at the same time scholars and researchers in the humanities have been approaching research in new computational methods such as data visualization, geographic information system (GIS) mapping, text mining, image analysis and topic modelling. Digital collections are presented as datasets that can be examined through computers or computer-aided tools and methods to facilitate digital humanities research.[3] To support this new viewpoint, the Institute of Museum and Library Services funded the 2016-2018 grant project *Always Already Computational: Collections as Data* that created a guide of standards and best practices for transforming digital collections to data for computational use and innovative research methods. The subsequent work in the 2018 Andrew W. Mellon funded

[1] Thomas G. Padilla, "Collections as data: Implications for enclosure," *College & Research Libraries News*, 79, no. 6 (2018): 296, https://doi.org/10.5860/crln.79.6.296.

[2] Florian Windhager et al., "Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges," *IEEE Transactions on Visualization and Computer Graphics*, 25, no. 6 (2019): 2311-2330, https://doi.org/10.1109/TVCG.2018.2830759; Monika Glowacka-Musial, "Visualization and Digital Collections," *Library Technology Reports*, 57, no. 1 (2021): 5-10, https://doi.org/10.5860/ltr.57n1.

[3] Rachel Wittmann, et al., "From Digital Library to Open Datasets: Embracing a "Collections as Data" Framework," *Information Technology & Libraries*, 38, no.4 (2019): 49–61, https://doi.org/10.6017/ITAL.V38I4.11101.

doi:10.1017/alj.2024.12

project *Collections as Data: Part to Whole* sought to promote responsible implementation and use of collections as data.[4] As part of the second phase grant, a working summit was conducted in April 2023 where 60 GLAM participants from 18 countries gathered in Vancouver at Internet Archive Canada for *Collections as Data: State of The Field and Future Directions*.[5]

Internationally, GLAM institutions have taken notice to the collections as data movement with many invested in developing and implementing practices to make collections available in machine-readable formats to enable and encourage computational research. The collections as data impetus in GLAM are largely due to the previously named two-part Collections as Data grant funded projects, and the larger OpenGlam movement, which encouraged and supported GLAM organizations like the Birmingham Museums Trust and Los Angeles County Museum of Art to make their collections open access without restrictions to users.[6]

Metadata, Open Data, Linked Data

The scope of data considered in the collections as data model is all encompassing. Trustworthy collections as data should include open, robust metadata, and under the care of stewards committed to their preservation. Examples of this data are images, descriptive metadata, optical character recognition (OCR) text, oral history texts, and full text transcriptions. Additionally, data that describe those data are in scope. For example, images and the metadata, finding aids, and/or catalogues that describe them are equally considered. Machine-readable open data is required to be made available under an open license and ideally published in a non-proprietary format like CSV, XML, JSON and enables use and reuse.[7]

Structured linked data is also presented as collections of data. It allows data from different sources to be connected and queried. Linked data requires two components – Uniform Resource Identifier (URI) and Resource Description Framework (RDF) structured data. Linked data uses identifiers to describe entities and enables relationships between entities, interoperability, and collaboration. Emerging standards for linked data provide new options of enhancing, contextualizing, linking, and reframing cultural heritage objects and collection data. Based on growing or emerging literature, Antonis Bikakis et al. posited, "Linked data and semantic web technologies are becoming increasingly important in creating, publishing, and analyzing cultural heritage data in digital humanities."[8] Furthermore, linked data provides a sustainable way of making digitised and born-digital archives more accessible, producing enhanced, integrated, and interoperable large-scale archival datasets available for reuse in multiple ways.[9]

Wikidata, the collaborative linked data repository behind Wikipedia and other Wikimedia projects, offers a low-barrier and high-result method for creating and using linked data in libraries and cultural heritage institutions. Many GLAM institutions, including the Art Institute of Chicago, the Museo Del Prado, and the University of Edinburgh, have shared their collections metadata and developed projects with Wikidata to build connections with Wikidata's rich data and wide audience.[10]

The Importance of Documentation

Making collections available in machine-readable formats is one step to facilitate computational research. Equally important are discussions on the processes and workflows required to turn collections into data and maintaining a critical perspective when handling the data.[11] A central approach for developing computational use of collections as data is shared documentation. It empowers others with direction to do the work and experiment, and "is a key element to foster data reuse by the community."[12] Collections as data work encompasses a variety of methods, including but not limited to text mining, computer vision, machine learning, artificial intelligence, data visualization, mapping, image analysis, and audio analysis. Many steps and decisions are involved about selection, description, formatting, data remediation, and delivery tools that provide open access and discovery to collection data. Examples of such documentation include workflows, application profiles, metadata schemas, data transformation actions, datasheets and codebooks.[13]

Documentation also cultivates transparency and collections as data integrity. To work transparently and sustain data integrity, digital cultural heritage (DCH) data stewards need to place the data in context and record data provenance, the relationships with the collections and communities, internal processes and

[4] Wittmann, et al., "From Digital Library to Open Datasets" 50; Thomas Padilla, "Always Already Computational," *Always Already Computational: Collections as Data*, 2018, https://collectionsasdata.github.io/; Thomas Padilla, "Part to Whole," *Collections as Data: Part to Whole*, 2019, https://collectionsasdata.github.io/part2whole/.

[5] "Collections as Data Futures: A Recap, A Resource, Next Steps," Collections as Data: Part to Whole, last modified May 4, 2023, https://collectionsasdata.github.io/part2whole/recap/.

[6] Sarah Ames and Stuart Lewis, "Disrupting the Library: Digital Scholarship and Big Data at the National Library of Scotland," *Big Data & Society* 7, no. 2 (2020): 1-6, https://doi.org/10.1177/2053951720970576; Foteini Valeonti, Melissa Terras, and Andrew Hudson-Smith, "How Open Is OpenGLAM? Identifying Barriers to Commercial and Non-Commercial Reuse of Digitised Art Images," *Journal of documentation* 76, no. 1 (2020): 1–26, https://doi.org/10.1108/JD-06-2019-0109.

[7] Thomas Padilla et al., "The Santa Barbara Statement on Collections as Data (V1)," n.d., https://collectionsasdata.github.io/statementv1/

[8] Antonis Bikakis, et al., "Editorial: special issue on semantic web for cultural heritage," *Semantic Web* 12, no. 2 (2021): 163–167, https://doi.org/10.3233/SW-210425.

[9] Ashleigh Hawkins, "Archives, linked data and the digital humanities: increasing access to digitised and born-digital archives via the semantic web," *Archival Science*, 22, (2022): 319–344, https://doi.org/10.1007/s10502-021-09381-0

[10] Will Kent, "Wikidata at the Art Institute of Chicago," Wikiedu, February 13, 2020, https://wikiedu.org/blog/2020/02/13/wikidata-at-the-art-institute-of-chicago/; Museo del Prado, "The Museo del Prado is once again at the forefront of museums online in its application of Artificial Intelligence to its collections," October 8, 2019, https://www.museodelprado.es/en/whats-on/new/the-museo-del-prado-is-once-again-at-the/33595ab1-92dd-460c-5831-4e68a042a21c; "Scottish Accused Witches: A project putting Scotland's accused witches on the map," Wikimedian

decision making with the collections, and transformations of collections or to its data. This enables institutions and data stewards to recognize and reduce biases, absences and uncertainty.[14] According to Sarah Ames, "reasoning behind how and why datasets have been produced, ensure a fuller understanding of the collections: this is an important aspect of contextualising libraries' digital and digitised collections." Communicating the provenance of cultural heritage data is vital to retain information from the original physical collection and allows users to know why, where, and how the data was collected and any modifications.[15]

Datasheets

Datasheets are one form of documentation that can prioritize transparency and accountability for DCH data and serve as a source for informed decisions about reusing datasets. According to Timnit Gebru, et al., datasheets for datasets can "mitigate unwanted societal biases in machine learning models, facilitate greater reproducibility of machine learning results, and help researchers and practitioners to select more appropriate datasets for their chosen tasks."[16] A datasheet accompanies and documents a dataset, detailing context required for reusing the data, such as its provenance, motivation, composition, collection process, and recommended uses. Although data source documents for DCH data can come in many forms for machines or humans to read, datasheets allow both a structured form to provide guidance on how to describe the datasets according to the reusers' information needs and a detailed narrative form whenever necessary. This flexible combination is especially advantageous for the diverse and intricate nature of DCH collections.[17]

Growing applications of various computational methods with large data in commercial and academic settings have led to mounting interest in machine learning (ML) with DCH data. In 2022, the Datasheets for Digital Cultural Heritage Working Group formed within the Europeana Research and EuropeanaTech Communities to modify the concept of ML datasheets for the cultural heritage sector. The group published their findings, recommendations, and a datasheet template proposal adapted for cultural heritage institutions in the *Journal of Open Humanities Data* in 2023.[18] Alkemade et al. assert that "DCH collections without proper information providing the required context needed for (re-)use, are prone to exploitation and misuse."[19]

DCH datasets differ from research datasets typically used in ML. Research datasets are typically gathered to create knowledge such as a a statistical sample to a specific research question. DCH data is rarely created for computation and they are presented with earlier constructed knowledge. DCH collections are a challenging concept as they are biased by definition and often built on preexisting collections created hundreds of years ago. Datasets can feature these multiple layers of selection and put forth a discriminatory viewpoint. The data originates from digitised items with associated text and metadata of physical heritage objects from curated collections. They are varied, possess specific characteristics, and include descriptive metadata from controlled vocabularies, ontologies, taxonomies or linked open data. DCH datasets will often have complex cultural elements, multilingualism, visual, audio, symbolic, and historical content. They can contain sensitive content and derogatory language, be copyright protected, and may grow or change over time. DCH stewards need to record such characteristics in datasheets to inform the reuse of datasets for academic and research purposes. Careful consideration is necessary for the intended and unintended use when applying ML or other computational methods to DCH datasets.[20] Alkemade et al. stress "the need for strong ethical commitments in the pursuit of responsible production, circulation, use and re-use of DCH datasets" and trust these commitments could be fostered by carefully documenting attributes in their datasheet template.[21]

The Checklist

Further adding to the spirit of shared documentation, the International GLAM Labs Community presented the 2022 webinar "Towards implementing Collections as Data in GLAM institutions" to inspire institutions to engage. The webinar presented real-life examples and a checklist that can be used for creating and evaluating digital collections for computational use.[22] A Checklist to Publish Collections as Data in GLAM Institutions was based on a literature review and

in Residence Use Cases, University of Edinburgh, accessed March 23, 2024, https://www.ed.ac.uk/information-services/help-consultancy/is-skills/wikimedia/wikidata/use-cases/scottish-witches.

[11] Sarah Ames, "Transparency, Provenance and Collections as Data: The National Library of Scotland's Data Foundry," *LIBER quarterly* 31, no. 1 (2021): 1–13, https://liberquarterly.eu/article/view/10880/11789; Henk Alkemade, et al., "Datasheets for Digital Cultural Heritage Datasets," *Journal of open humanities data* 9 (2023): 1–11, https://doi.org/10.5334/johd.124.

[12] Thomas Padilla, et al., "50 Things—always already computational: collections as data," Zenodo, May 20, 2019, https://doi.org/10.5281/zenodo.3066237.

[13] Thomas Padilla, et al., "Vancouver Statement on Collections as Data," Zenodo, September 13, 2023, https://doi.org/10.5281/zenodo.8342171.

[14] Ames, "Transparency, Provenance and Collections as Data," 10; Padilla, et al., "Vancouver Statement," 4.

[15] Ames, "Transparency, Provenance and Collections as Data," 10.

[16] Timnit Gebru et al., "Datasheets for Datasets," *Communications of the ACM*, 64, no. 12 (2021): 86–92, https://doi.org/10.1145/3458723.

[17] Alkemade, et al., "Datasheets for Digital Cultural Heritage Datasets," 2.

[18] "Datasheets for digital cultural heritage Working Group," *Europeana Pro,* accessed April 4, 2024, https://pro.europeana.eu/project/datasheets-for-digital-cultural-heritage-working-group.

[19] Alkemade, et al., "Datasheets for Digital Cultural Heritage Datasets," 2.

[20] Alkemade, et al., "Datasheets for Digital Cultural Heritage Datasets," 3-4; Catherine Nicole Coleman, "Managing Bias When Library Collections Become Data," *International Journal of Librarianship,* 5, no. 1 (2020), 8-19, https://doi.org/10.23974/ijol.2020.vol5.1.162

[21] Alkemade, et al., "Datasheets for Digital Cultural Heritage Datasets," 5.

[22] "Webinar: Towards implementing Collections as Data in GLAM," *International GLAM Labs Community*, accessed March 23,

feedback from the community and offers institutions a way to adopt collections as data principles into workflows following best practices. In 2023, Gustavo Candela, et al. published the context surrounding this work and "an extensive demonstration of how to make available digital collections suitable for computational use, giving particular attention to data quality, planning and experimentation" in *Global Knowledge, Memory and Communication.*[23]

Community and Collaborations

Enhancing the landscape are transnational and national community organizations, consortia, and partnerships advancing to support, educate, and promote research projects and resources in machine learning, collections data, and the digital humanities as strategic priorities. The following are a few examples.

- In 2018, *DARIAH Beyond Europe: Collections as Data: Digital Collections for Emerging Research Methods* took place at the Library of Congress. The collaboration was a series of workshops joining European Digital Research Infrastructure for Arts and Humanities (DARIAH) community and Digital Arts and Humanities projects in the US and Australia to exchange ideas and discuss how their digital humanities initiatives and approaches could intersect in the North American academic community.[24]
- Formed in 2018, the International GLAM Labs Community supports and encourages GLAM institutions to publish their collections as data and includes 250 members, from more than 60 institutions, in over 30 countries. Most notably, GLAM Labs have produced innovative initiatives that reuse digital collections with computational methods by means of Jupyter Notebook, including institutions Biblioteca Virtual Miguel de Cervantes Labs, the British Library, National Library of Scotland, and the Library of Congress. A free and open-source software, Jupyter Notebook is a web application used to create and share live code, visualizations, images, and text within a single web page and facilitates easy documentation of data cleaning, analysis, and exploration. The tool makes it possible to read collection data, but also download, analyse and visualise data within a browser. Additionally, the GLAM Workbench website created by Tim Sherratt, is devoted to sharing Jupyter notebook examples, tools, and tutorials to assist in the work with collections data from GLAM institutions.[25]
- The UK Arts and Humanities Research Council (AHRC) and USA National Endowment for the Humanities (NEH) created the NEH/AHRC New Directions for Digital Scholarship in Cultural Institutions to encourage UK and US collaboration to advance machine learning work. The 2021 and 2022 program awarded a total of twenty research projects.[26]
- *Cultural Heritage Data as Humanities Research Data* was the theme for the DARIAH-EU annual event in June 2023. 200 participants from over 30 countries attended the three-day conference and explored questions such as "What does it mean for cultural heritage institutions to provide access to their 'collections as data'?" and "Can we think of a humanities research data continuum?"[27]
- The Royal Library of Belgium and the Europeana Foundation in February 2024 offered the collaborative workshop *Collections as Data: collaborating across data spaces for cultural heritage and open science.* Europeana Research is active with aligning the collections as data international movement principles with the European data space for cultural heritage.[28]

Discovery, Inspiration and Innovation: Projects and Initiatives

Treating collections in cultural heritage institutions as data encourages novel approaches to the use of collections and generates new perspectives and knowledge. Following are case studies that exhibit innovative and inspiring practices and discovery into new perspectives and narratives.

In 2011, Rijksmuseum was the first museum to offer free and unrestricted access to thousands of high-quality images of artworks in its collection that were in the public domain, allowing anyone to download and use these images without any limitations. Several museums followed its lead, including the Metropolitan Museum of Art and the Statens Museum Kunst, with more institutions participating each year. An institution's commitment to open access requires a continuous

2024, https://glamlabs.io/webinar-towards-implementing-collections-data-glam/.

[23] Gustavo Candela, et al., "A checklist to publish collections as data in GLAM institutions," *Global Knowledge, Memory and Communication*, ahead-of-print, no. ahead-of-print (2023), https://doi.org/10.1108/GKMC-06-2023-0195

[24] *DARIAH Beyond Europe: Collections as Data: Digital Collections for Emerging Research Methods*, 2018, Video, https://www.loc.gov/item/webcast-8749/.

[25] International GLAM Labs Community, accessed March 23, 2024, https://glamlabs.io/; GLAM Workbench, accessed March 23, 2024, https://glam-workbench.net/.

[26] National Endowment for the Humanities, "New Grants Awarded by NEH and UK Arts and Humanities Research Council to Support Digital Innovation at Cultural Institutions," January 12, 2022, https://www.neh.gov/news/new-grants-awarded-neh-and-uk-arts-and-humanities-research-council-support-digital-innovation; NEH/AHRC New Directions for Digital Scholarship in Cultural Institutions, *National Endowment for the Humanities*, accessed March 23, 2024, https://www.neh.gov/divisions/odh/new-directions.

[27] Amelia McConville, "Recap of the Annual Event 2023: Cultural Heritage Data as Humanities Research Data?" *DARIAH-EU*, July 7, 2023, https://www.dariah.eu/2023/07/07/recap-of-the-annual-event-2023-cultural-heritage-data-as-humanities-research-data.

[28] "Collections as Data: collaborating across data spaces for cultural heritage and open science," Europeana Pro, last modified February 12, 2024, https://pro.europeana.eu/event/collections-as-data-collaborating-across-data-spaces-for-cultural-heritage-and-open-science.

investment of time, expertise, and dedicated work. Many provide digital images of objects from the collection, descriptive metadata, and bibliographic data without restrictions on reuse with the Creative Commons Zero (CC0) designation.[29]

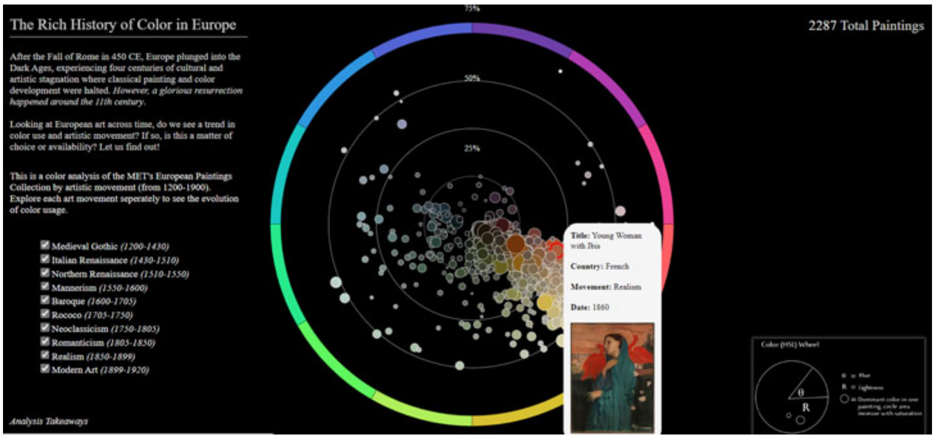
As a commitment to open access, many GLAM institutions now utilize application programming interfaces (APIs) to openly share their collections and its data. The Rijksmuseum, the Victoria and Albert Museum, the Smithsonian, Harvard Art Museums, the Library of Congress, and Europeana are among the many embracing APIs. An API is a standard way software applications interact and exchange data. It is tool that allow computers to read and analyse a changing set of information and a proficient way to communicate with databases on remote servers to return specific information, enabling individuals to create new experiences.[30]

Partnering with the Metropolitan Museum of Art (The Met), students at the Parsons' School of Art design, visualize and interpret the museum's open access collection data using its API. The Met's API offers access to The Met's Open Access data and to corresponding high resolution images that are in the public domain. Students created projects that showed objects in novel ways, revealing new connections and narratives. *The Met Ceramics Lookbook* uses the API to compare the textures of ceramic objects in the collection. *Rich History of Color in Europe* (Figure 1) looks at the evolution of colour usage in the European Paintings collection across time and artistic movements. *The Landscape Generator* is an interactive website made to encourage the user to engage with the collections' landscape paintings in a unique way, leading them to notice details they may miss with a traditional view.[31]

Along with open access of images and data, many leading cultural institutions support the International Image Interoperability Framework (IIIF). IIIF is a set of open standards for delivering high-quality, attributed digital objects online at scale and is the sharing technology behind many innovative digital humanities projects. A fun and novel application of IIIF is taking artworks out of museums and into the popular Nintendo video game Animal Crossing with the Animal Crossing Art Generator tool created by the Getty. All collections using IIIF can be applied.[32]

Computational techniques applied to open data can provide insight into collections. Using standard computer vision methods to undertake research into the history of printmaking at the University of Oxford, a collaboration with the faculty of English literature, Bodleian Libraries, and the Department of Engineering Science was able to identify woodcuts reused in different publications and dated and ordered the publications by finding tiny differences in the condition of the woodcuts (Figure 2).[33]

Following guidance from Padilla's "On a Collections as Data Imperative," the University of Utah converted five digital collections into datasets in their pilot for a collections as data strategy. Datasets included geographic coordinates and mapping, genealogical information from newspaper obituaries, transcription of historical records, and mining-related oral history texts. They tested the data with various digital humanities methods for computational exploration including topic modelling, a type of text analysis that clusters word groups or potential topics on



The Rich History of Color in Europe

After the Fall of Rome in 450 CE, Europe plunged into the Dark Ages, experiencing four centuries of cultural and artistic stagnation where classical painting and color development were halted. *However, a glorious resurrection happened around the 11th century.*

Looking at European art across time, do we see a trend in color use and artistic movement? If so, is this a matter of choice or availability? Let us find out!

This is a color analysis of the MET's European Paintings Collection by artistic movement (from 1200-1900). Explore each art movement seperately to see the evolution of color usage.

- ☑ Medieval Gothic *(1200-1430)*
- ☑ Italian Renaissance *(1430-1510)*
- ☑ Northern Renaissance *(1510-1550)*
- ☑ Mannerism *(1550-1600)*
- ☑ Baroque *(1600-1705)*
- ☑ Rococo *(1705-1750)*
- ☑ Neoclassicism *(1750-1805)*
- ☑ Romanticism *(1805-1850)*
- ☑ Realism *(1850-1899)*
- ☑ Modern Art *(1899-1920)*

Analysis Takeaways

2287 Total Paintings

Title: Young Woman with Ibis

Country: French

Movement: Realism

Date: 1860

75%
50%
25%
Color (HSI) Wheel
θ
R
θ = Hue
R = Lightness
○ = Dominant color in one painting, circle area increase with saturation

Fig. 1. Nour Zein's *Rich History of Color in Europe* is a colour analysis of The Met's European Paintings Collection by artistic movement, https://nourzein.github.io/Major-Studio1/mobile/.

[29] Valeonti, Terras, and Hudson-Smith, "How Open Is OpenGLAM?" 1; Gustavo Candela, et al., "Reusing digital collections from GLAM institutions," *Journal of Information Science*, 48, no.2 (2022), 251–267, https://doi.org/10.1177/0165551520950246.

[30] Bill Doerrfeld, "How Museums Are Using APIs to Inspire Art Lovers Worldwide," *Nordic APIs*, August 3, 2020, https://nordicapis.com/how-museums-are-using-apis-to-inspire-art-lovers-worldwide/; Rachel Kraus, "A Museum Without Walls: How the Met is Bringing Its Ancient Collection Online," *Mashable*, October 25, 2018, https://mashable.com/article/the-met-museum-api#Lti3vVoahsqi.

[31] "The Metropolitan Museum Partnership 2019," *New School MS Data Visualization*, accessed March 23, 2024, https://parsons.nyc/met-museum-2019/;Benjamin Korman and Maria Kessler, "Show Your Work: Parsons Students Design Stunning Data Visualizations with Met Open Access API," *Metropolitan Museum of Art*, February 7, 2020, https://www.metmuseum.org/articles/met-api-parsons-data-visualization; Zui Chen, Met Ceramics Lookbook, accessed March 25, 2024, https://azuic.github.io/the-met-ceramics-lookbook/; Amanda Anderson-You, Landscape Generator, accessed March 25, 2024, https://github.com/amandersonyou/MajorStudio1_AAY/blob/master/Met_Interactive/README.md; Nour Zein, Rich History of Color in Europe, accessed March 25, 2024, https://nourzein.github.io/Major-Studio1/mobile/.

[32] "Animal Crossing art generator," *Getty*, accessed March 20, 2024, https://experiments.getty.edu/ac-art-generator; "How to Build an Art Museum in Animal Crossing," *Getty*, accessed March 20, 2024, https://www.getty.edu/news/how-to-build-an-art-museum-in-animal-crossing/

[33] Joon Son Chung, et al., "Re-presentations of Art Collections," in *Computer Vision - ECCV 2014 Workshops. ECCV 2014, Lecture Notes in Computer Science*, ed. Lourdes Agapito, Michael M. Bronstein, Carsten Rother (Cham, Switzerland: Springer International Publishing,

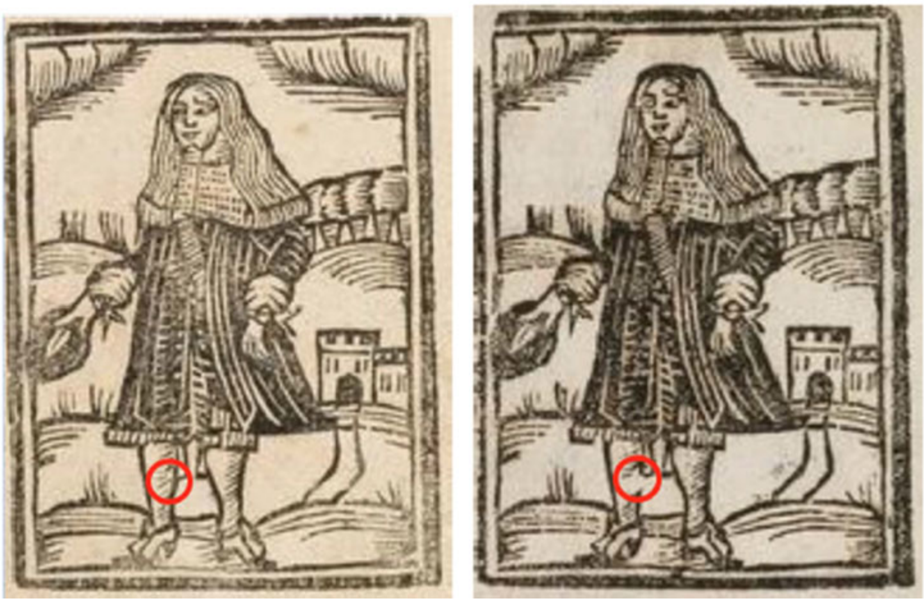


Fig. 2. Woodcut illustrations printed from the same block with tiny differences due to wear and tear. With standard computer vision methods, detecting such differences can aid temporal ordering of prints and publications. The University of Oxford project is discussed in Joon Son Chung, et al. "Re-presentations of Art Collections," https://doi.org/10.1007/978-3-319-16178-5_6

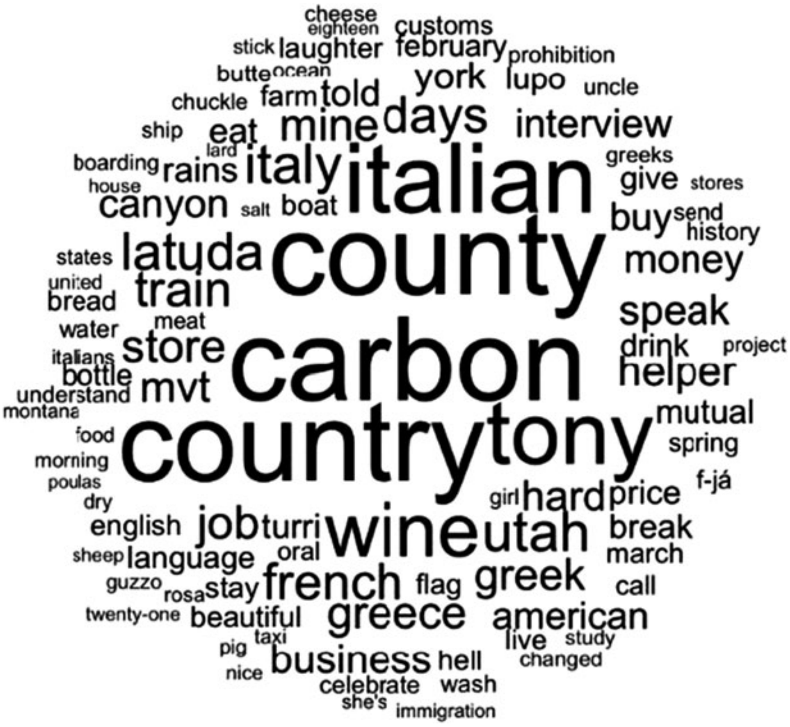


Fig. 3. Word cloud of topic model for the mining-related oral histories collections at University of Utah's Digital Library. The project is discussed in Rachel Wittmann, et al. "From Digital Library to Open Datasets: Embracing a "Collections as Data" Framework," https://doi.org/10.6017/ITAL.V38I4.11101

a set of documents. Word cloud outputs of the mining oral history texts revealed topics not in the collections' descriptive metadata (Figure 3). Topic modelling allows for the discoverability of valuable topics that may be hidden in the content of vast documents.[34]

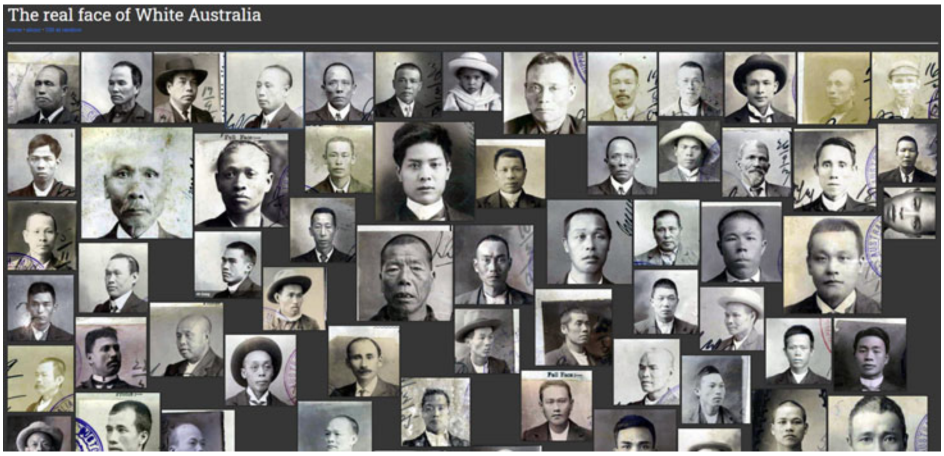
When working with collections as data, Padilla reminds practitioners that "collections as data represent lived experience and we should respect that."[35] Ethical concerns are integral to collections as data. The scale of some collections may exclude narratives and institutions should be aware of existing gaps and attempt to counter and work against their repetition.[36] The following projects take the opportunity to counter dominant historical narratives.

Kate Bagnall and Tim Sherratt created *The Real Face of White Australia* (Figure 4) by employing the power of facial detection script to extract faces from archival government documents to reveal the falsehood of Australia's early twentieth self-definition as a white man's country. The project humanizes the historical data and brings to light the country's lesser-known history of discriminatory and racist policies toward non-white populations to deny them as Australians.[37]

Similarly, Illinois State University's *Agency through Otherness: Portraits of Performers in Circus Route Books, 1875-1925*, humanizes data found in the circus route books collection and elucidates the disparities among people during this period. The project features interactive map visualizations of circus routes with shared open data of historical railroads, population numbers, and US Native Lands. The "Routing the Circus 1875-1925" map (Figure 5) showcases circus stops alongside population and railroads data illustrating how the circus's dominant narratives of non-Western cultures as human curiosities spread across the country, reinforcing colonialist notions of power and racist hierarchy. The "Native Lands and the Wild West Show" map (Figure 6) juxtaposes circus routes with the genocide and forced relocation of Native communities to use the circus routes history to underscore these atrocities.[38]

According to Windhager et al., "Datasets of cultural heritage collections are connected to historical contextual information and knowledge. Linked data can help with this contextualization in visualizing and interpreting collections in broader cultural and societal spaces."[39] The following projects provide contextualization and broader cultural interpretations with linked data.

The Art Tracks Provenance Project by the Carnegie Museum of Art used provenance data to create map visualizations to show the movement of a painting over time (Figure 7). The combined efforts of technologists, curators, and provenance researchers worked to turn the history of ownership of an object into structured data. They designed a structured provenance data model and used linked data to supply the events and topics surrounding an art piece to offer more context and meaning.[40]



The real face of White Australia

Fig. 4. Kate Bagnall and Tim Sherratt, *The Real Face of White Australia*. The creators extracted faces from archival government documents data using a facial detection script to counter the historical narrative of a "White Australia," https://www.realfaceofwhiteaustralia.net/.

2015), 85-100, https://doi.org/10.1007/978-3-319-16178-5_6.

[34] Wittmann, et al., "From Digital Library to Open Datasets," 51-52.

[35] Padilla, "Collections as Data: Conditions of Possibility," *Medium*, Oct 28, 2016, https://tgpadillajr.medium.com/collections-as-data-conditions-of-possibility-494805bf16be.

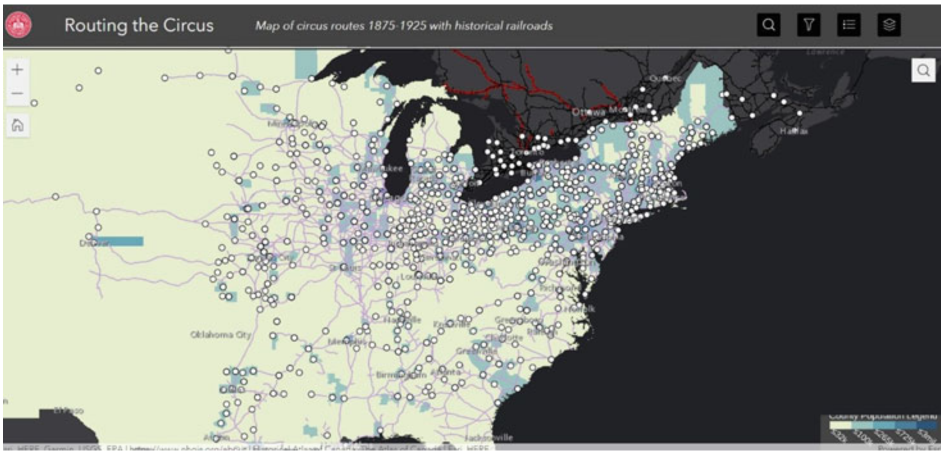
[36] Padilla, et al., "Vancouver Statement," 2.

[37] Padilla, "Collections as Data: Conditions of Possibility."; Kate Bagnall and Tim Sherratt, *The Real Face of White Australia*, accessed March 20, 2024, https://www.realfaceofwhiteaustralia.net/.

[38] "Routing the Circus," in Agency Through Otherness: Portraits of Performers in Circus Route Books, 1875-1925, ed. Angela Yon, accessed April 3, 2024, https://scalar.usc.edu/works/circus-route-books-project/routing-the-circus; Mariah Wahl, "Native Performance and Identity in The Wild West Show," in Agency Through Otherness: Portraits of Performers in Circus Route Books, 1875-1925, ed. Angela Yon, accessed April 3, 2024, https://scalar.usc.edu/works/circus-route-books-project/native-performance-in-the-wild-west.

[39] Windhager et al., "Visualization of Cultural Heritage Collection Data," 2312.

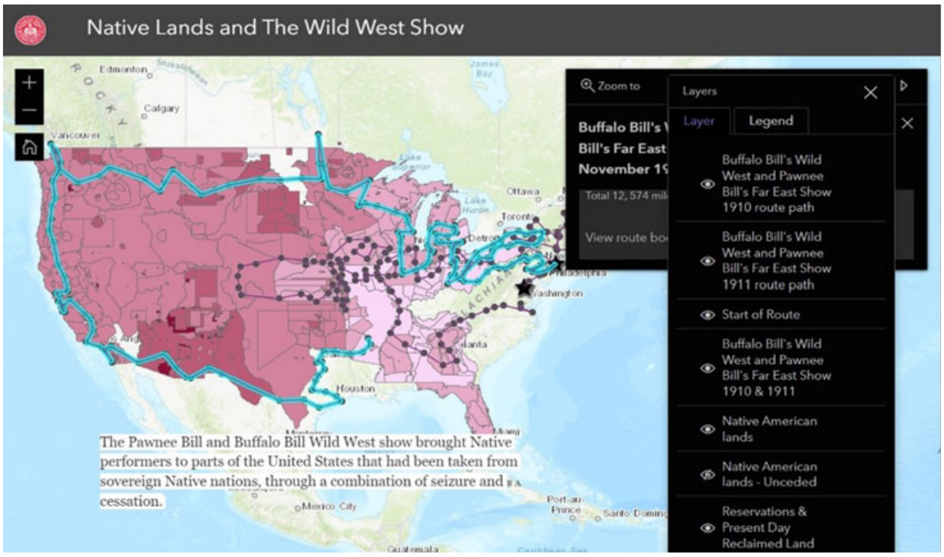
[40] David Newbury, "Art Tracks: a technical deep dive," (presentation slides, 2016 Digital Provenance Symposium, Carnegie Museum of Art, Pittsburgh, PA, October 14, 2016); "The Northbrook Collection as Case Study," *Northbrook Provenance Carnegie Museum of Art*, accessed March 18, 2024, https://northbrook.cmoa.org/about/case-study/#digital-provenance-symposia; "2016 Digital Provenance Symposium," *Art Tracks a project at Carnegie Museum of Art,* accessed March 18, 2024, https://www.museumprovenance.org/pages/scholars_day_2016/.



Routing the Circus

Map of circus routes 1875-1925 with historical railroads

Fig. 5. “Routing the Circus 1875-1925” map, circus route stops overlaid with historical railroad lines and county population census data illustrate how the circus’s dominant narratives of colonialist notions of power and racist hierarchy spread across the United States, https://scalar.usc.edu/works/circus-route-books-project/routing-the-circus.



Native Lands and The Wild West Show

The Pawnee Bill and Buffalo Bill Wild West show brought Native performers to parts of the United States that had been taken from sovereign Native nations, through a combination of seizure and cessation.

Buffalo Bill's Bill's Far East November 19

Total 12, 574 mil

View route bo

Layers

Layer | Legend

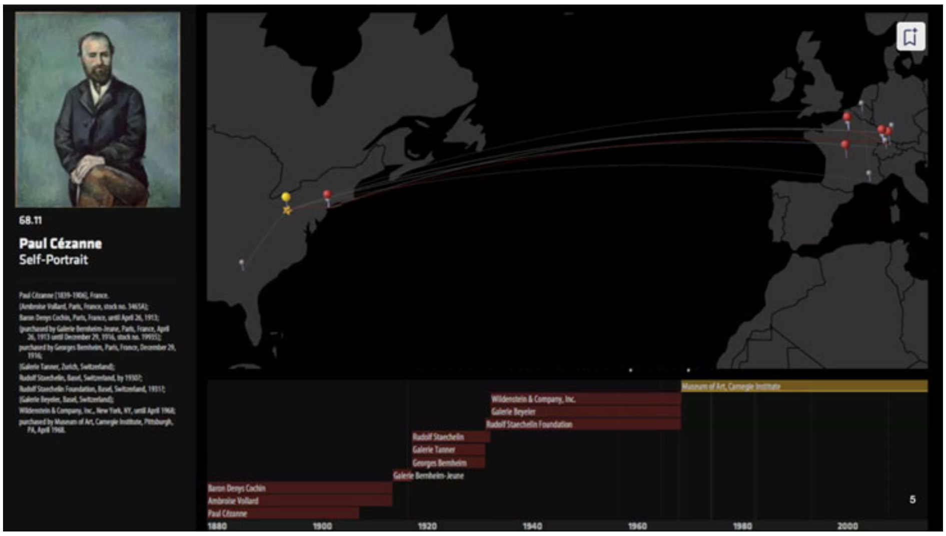
- Buffalo Bill's Wild West and Pawnee Bill's Far East Show 1910 route path
- Buffalo Bill's Wild West and Pawnee Bill's Far East Show 1911 route path
- Start of Route
- Buffalo Bill's Wild West and Pawnee Bill's Far East Show 1910 & 1911
- Native American lands
- Native American lands - Unceded
- Reservations & Present Day Reclaimed Land

Fig. 6. “Native Lands and the Wild West Show” map, Buffalo Bill’s Wild West and Pawnee Bill’s Far East Show 1910 and 1911 route paths overlaid with Native American lands ceded and unceded underscore the genocide and forced removal of Native communities with circus routes history, https://scalar.usc.edu/works/circus-route-books-project/native-performance-in-the-wild-west.

An international initiative, the Program for Cooperative Cataloging (PCC) - Wikidata Pilot in September 2020, drew over seventy academic and cultural organizations and librarians in many roles to explore identity management and linked data through Wikidata. Several contributing institutions focused on art-related linked data projects, including the Frick Art Reference Library, New Mexico State Library, the Harry Ransom Center, and the National Gallery of Art in Washington, DC.[41]

Wikidata also works within Wikibase, which is an open-source software that allows users to generate a custom and structured data repository. One notable Wikibase repository is *Enslaved: Peoples of the Historical Slave Trade*. The site, which launched in December 2020, is a collaborative effort funded by the Andrew W. Mellon foundation and features research from scholars at Michigan State University, Harvard Hutchins Center for African & African American Research, University of Maryland, and other institutions. Visitors can target searches

[41] Program for Cooperative Cataloging, “Wikidata:Wikiproject PCC Wikidata Pilot,” accessed March 29, 2024, https://www.wikidata.org/wiki/Wikidata:WikiProject_PCC_Wikidata_Pilot; New Mexico State Library, “Wikiproject PCC Wikidata Pilot/ New Mexico State Library Projects,” Wikidata, accessed March 29, 2024, https://www.wikidata.org/wiki/Wikidata:WikiProject_PCC_Wikidata_Pilot/New_Mexico_State_Library_Projects/NM_Artworks; Frick Art Reference Library, “Wikiproject PCC Wikidata Pilot/ Frick Art Reference Library,” Wikidata, accessed March 29, 2024, https://www.wikidata.org/wiki/Wikidata:WikiProject_PCC_Wikidata_Pilot/Frick_Art_Reference_Library;Harry Ransom Center, “Wikiproject PCC Wikidata Pilot/ Harry Ransom Center,” Wikidata, accessed March 29, 2024, https://www.wikidata.org/wiki/Wikidata:WikiProject_PCC_Wikidata_Pilot/Harry_Ransom_Center; National Gallery of Art Library, “Wikiproject PCC Wikidata Pilot/ National Gallery of Art Library,” Wikidata, accessed March 29, 2024, https://www.wikidata.org/wiki/Wikidata:WikiProject_PCC_Wikidata_Pilot/National_Gallery_of_Art_Library



68.11

Paul Cézanne
Self-Portrait

Paul Cézanne (1839-1906), France.
(Ambroise Vollard, Paris, France, stock no. 3465A);
Baron Denys Cochin, Paris, France, until April 26, 1913;
(purchased by Galerie Bernheim-Jeune, Paris, France, April 26, 1913 until December 29, 1916, stock no. 19935);
purchased by Georges Bernheim, Paris, France, December 29, 1916;
(Galerie Tanner, Zurich, Switzerland);
Rudolf Staechelin, Basel, Switzerland, by 1930?;
Rudolf Staechelin Foundation, Basel, Switzerland, 1931?;
(Galerie Beyeler, Basel, Switzerland);
Wildenstein & Company, Inc., New York, NY, until April 1968;
purchased by Museum of Art, Carnegie Institute, Pittsburgh, PA, April 1968.

Fig. 7. Map visualization with timeline showing the movement of a painting over time created with provenance data and linked data technologies by the Carnegie Museum of Art Tracks Provenance Project. David Newbury discussed the method in "Art Tracks: a technical deep dive," at the 2016 Digital Provenance Symposium," *at Carnegie Museum of Art,* https://www.museumprovenance.org/pages/scholars_day_2016/.

through more than 5 million data points, including names, genders, birthplaces, places of death, and occupations of individuals collected from existing research and shared datasets conducted by multiple institutions around the world; made possible by linked data technologies. The data comes from census reports, baptismal, shipping, and sales records that tell the story of people captured by the transatlantic slave trade, including in-depth biographies. The entries run from the 15th to the late 19th centuries and span Western Europe, Africa, and North and South America.[42]

Challenges and Considerations

Collections as data is compelling work, yet there are considerations to ponder. Institutions face challenges when resources are scarce, and methods to adopt collections as data vary and guidelines are not clear. Practitioners find there is a gap between theory and best practices in preparing the data efficiently and effectively to serve researchers.[43] Listed are some common challenges:

Scalability

Scaling up work is a requirement for experimentation and often many persons and different departments need to be involved in collections as data projects work. There is a diversity of roles and skills necessary to facilitate collections as data work. Creating and maintaining working relationships between individuals and organizational units are an integral part for success. Without grant funding for collections as data projects, existing resources and infrastructure may not be able to support efforts into regular day-to-day processes, workflows and produce an open data delivery platform.[44]

Rights and restrictions

Copyright, privacy, and confidentiality concerns exist, and obligations to the collection donor or creator need to be considered. There are ethical issues on the control of the data, terms of service requirements, awareness of potential uses of data and, threat of possible reidentification from sharing data.[45]

Researcher needs

A greater understanding of researcher needs is imperative. How should librarians invest for unknown future users with as of yet unknown needs? Creating

[42] Enslaved Peoples of the Historical Slave Trade, accessed March 20, 2024, https://enslaved.org/; Cogan Shimizu, et al., "The Wikibase Approach to the Enslaved.Org Hub Knowledge Graph," in: *The Semantic Web – ISWC 2023, ISWC 2023, Lecture Notes in Computer Science, vol 14266*, ed. Terry R. Payne, et al. (Cham: Springer International Publishing, 2023), 419–434, https://doi.org/10.1007/978-3-031-47243-5_23

[43] Candela, et al., "A checklist to publish collections as data in GLAM institutions," 3; Cory Lampert and Emily Lapworth, "What do we mean by "Collections As Data" (CAD)?" *UNLV University Libraries.* https://www.library.unlv.edu/whats-new-special-collections/2020/2020-03/what-do-we-mean-collections-data-cad-cory-lampert-emily.

[44] Ames and Lewis, "Disrupting the Library," 4; Weber, Chela Scott, "Collections as Data: Nascent progress and common need," *Hanging Together the OCLC Research Blog*, April 20, 2023, https://hangingtogether.org/collections-as-data-nascent-progress-and-common-need/.

[45] Candela, et al., "A checklist to publish collections as data in GLAM institutions," 3; Weber, "Collections as Data."

collections as data requires a level of educated speculation as to what researchers will want to access, what metadata fields they will be interested in manipulating, and in what formats they will need their data. Should digital collections be converted into data in anticipation of use or should collections be converted by request?[46]

Gaps in knowledge

Gaps and biases exist in datasets and care is needed to evaluate data and their potential harm. Education is required on the importance of contributing knowledge to strengthen marginalized voices and safeguard against the further silencing of populations.[47]

Conclusion

The blossoming collections as data movement is gaining momentum among GLAM and institutions are beginning to publish datasets for reuse in creative and inspiring ways. Strategic initiatives, experimentation, innovation, and inspirational learning are occurring as the fields of digital libraries and digital humanities progress and work to develop sustainable approaches for collections as data programs. Within the past decade public initiatives, grant funded projects, and emerging experiments reveal increased community implementation of collections as data and pioneering possibilities into the future. Institutions of all sizes collectively and individually have invested in developing and supporting computational use of collections as data. The landscape of open access, open data, and linked data enables new perspectives and knowledge, exposure of lesser-known narratives, and innovative exploration of critical questions in the humanities. In spite of advancing initiatives, the collections as data paradigm is still evolving. There are future opportunities to work collectively to address challenges and shared needs in the adoption of collections as data for computational, non-traditional, and innovative research. These are critical areas as the GLAM community faces information provision in a world of large-scale data, algorithms, and artificial intelligence.

Angela Yon
Assistant Professor/Cataloging & Metadata Librarian
Illinois State University
Milner Library
Normal, Illinois 61761
USA
Email: ayon@ilstu.edu

[46] Wittmann, et al., “From Digital Library to Open Datasets” 59; Weber, “Collections as Data.”

[47] Coleman, “Managing Bias,” 10-12; Padilla, et al., “Vancouver Statement, 2.”