Accelerating Precision Research and Resolution Through Computational Archival Science Pedagogy

Sarah A. Buchanan
University of Missouri
Columbia, USA
buchanans@missouri.edu

Jennifer L. Wachtel
University of Maryland and National Archives and Records Administration ◊
Washington, D.C., USA
jw23@terpmail.umd.edu

Jennifer A. Stevenson
Defense Threat Reduction Agency
Fort Belvoir, USA
jennifer.a.stevenson12.civ@mail.mil

2023 IEEE International Conference on Big Data (BigData) | 979-8-3503-2445-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/BigData59044.2023.10386566

***Abstract*—Use of archival collections is accelerated by the presence of finding aids, which communicate the arrangement and description of collection contents. To arrive at the optimal arrangement of a collection, archivists rely on some item-level processing or knowledge gained by exploring and manipulating digital reproductions of the contents. In this paper we consider archival student and instructor perspectives from hands-on course experiences directly with two distinct collections: one pertaining to the development, 2017 transfer and launch, and ongoing maintenance of the International Research Portal for Records Related to Nazi-Era Cultural Property (IRP2), and one a selection of unclassified catalog entries about digitized nuclear science reports. Visualizing is a data practice that permits the discovery of key content patterns, identification of computational models to be carried out to aid further analysis, and query-resolution for subject experts with precise – and historically significant – research questions. While archival data visualizations have previously been implemented as an extension of descriptive work including finding aid element counts, here we connect visualization to the work of archival outreach and access. We study how visualizations generated by groups of students working with textual and numerical dataset portions can ultimately accelerate time-sensitive uses of collections.**

Keywords—computational thinking, information science education, information visualization, provenance research, data transformation

I. Introduction

Archival collections are characterized by hierarchical levels of organization. Known-items, which are the most familiar target of research by subject experts, may be comprised of multiple documents or pages. In turn that item nests under a file unit, which nests under a series, which nests under a record group and/or subgroup, which finally is one of many housed in a depository [1]. Such a framework informs the archival tasks of arrangement and processing, culminating in public access or awareness of the materials. Collection management, among other topics in the realms of teaching, recruitment to the profession, and professional development, was highlighted by the Joint Committee on Archives/Libraries/Museums (CALM) as a key area where prospective, newly established, and experienced archivists alike can continue to make valuable investments of resources in order to enhance their digital presence. The CALM, before it was disbanded in late 2017, organized a trio of sessions at the major library, archival, and state and local history conferences that year which critically addressed work ongoing in heritage institutions to become more inclusive environments [2]. Crucially, collections work also happens to be among the most attractive and desirable skills for students in training to be archivists. Processing experience – an umbrella term inclusive of arrangement, description, digitization and other areas – becomes an activity sufficiently worthwhile of support in an archival curriculum, where it will reinforce the professional orientation of the program as students progress towards graduation. Our focus in this paper is to explore collection / data visualization and analysis practices that have been carried out in graduate archival education settings, in pursuit of an ethos of building computational skills and computational thinking approaches during the student experience. This paper uses names, initials, or a pseudonym to attribute specific material based on those individuals' stated preferences.

II. Literature on Collection Visualization: Three Roles

Visualization serves at least three functional roles in computational archival work, which we outline here. Apart from archival collection uses, data visualization helps thematize and characterize large datasets for broader outreach and access purposes. When they are released, public records datasets are nearly always in "rough shape" (NL, graduate class guest lecture with data journalist, September 9, 2019) due to the human (role) transfers or format transfers involved during their (by-) production. They require the use of pre-visualization tools like OpenRefine so communicators can make effective and interactive multimedia / channel / platform presentations of information contained in the datasets. Journalists on a national politics or elections beat for example may take an interest in fundraising data for political campaigns in a given state: "those are all entered through data entry software from physical pieces of paper, and there are a lot of interesting messes there" (NL, September 9, 2019). Data provenance disclosure can promote trust and transparency, countering a tenor of mistrust toward institutions with more humane ways for people – especially those who are subjects in the records – to DIY and participate if not improve upon the work of algorithmic / scientific reproducibility [3]. Once ascertained and/or disclosed, provenance serves a variety of purposes limited only by the researcher's end-goals including recall, replication, action

◊ The ideas expressed in this paper are entirely our own and reflect neither those of the National Archives and Records Administration or any other federal agency, nor of the University of Maryland where the work occurred.

This work was supported by the Institute of Museum and Library Services, grant RE-252287-OLS-22.

recovery, collaborative communication, presentation, and meta-analysis [4]. Wrisley [5] underscores "the necessity, and challenges, of a collaborative modeling and design process" between information experts and subject experts working together. The framework that communications scholars Matei and Hunter [6, p. 315] provide for data journalism views science itself as a "quest to provide unexpected answers to questions about observable realities" and valorizes simple chart visualizations, while Saffran's et al. [7, p. 12] interdisciplinary team stresses the value of authenticity and "the openness of the communicator toward revision – both past and future" to build public trust and credibility around any information presentation.

Though our work below connects most closely to data analysis tasks and systems thinking practices, visualization techniques are also implemented to support archival description and especially those descriptive products such as finding aids that may be harvested as part of aggregation efforts. The project called Building a National Finding Aid Network (NAFAN) [8] reconciles with issues around such aggregators' aging infrastructure, including the wide lack of EAD3 implementation, and the need to implement a shared governance structure for ensuring long-term access to the finding aids that have been so meticulously produced by archival workers across the country. Visualization will be explored in our paper as a complement to the arrangement of an archival collection: very often a concurrent step with description and, when taken together, constituting the work of transformative processing.

A. A Data Practice

In contrast to search, visualization of archival collections promises a more expansive and less constraining means of introducing a user to cultural heritage materials of interest. Importantly, visualization can be thought of as part of a suite of activities recognizing that users may not have a known result in mind, but are turning to various information tools to browse possibilities, satisfy a "leisurely curiosity" [9], or experience serendipity. Even so, libraries and archives have not yet fully integrated data visualization practices – with all of their attendant ethical, instructional, and source-dependent considerations – into professional praxis as deeply as have fields like new media studies and data journalism, though a cohort of academic fellows is making significant inroads in that respect [10]. Data visualization can support the broader project of information science to promulgate everyday strategies for critical thinking and decision-making [11] and engaging responsibly in civic society.

B. A Humanized Story of Quantitative Collections Data

The relatively recent approval of *Guidelines for Standardized Holdings Counts and Measures for Archival Repositories and Special Collections Libraries* by the two major American archival and research library organizations in 2019 is a significant development in the profession's ongoing effort to clearly and effectively communicate archival content information to general audiences [12]. The result of five years of dedicated multidisciplinary analysis of quantitative methods and techniques in use in collecting institutions, the *Guidelines* articulate both recommended and optional counting mechanisms for intellectual units, physical space, and digital space occupied. ArchivesZ offers proof of concept of some intriguing research questions that can be opened by displaying attributes such as collection sizes, temporality, and geographical heat maps either within or across repositories [13]. Given that archivists seize upon chances to meet users and assert the unique and singular qualities of their collections, agreement on basic measurement units now grants archivists significant freedom to communicate such information graphically. Its deceptively simple accessibility underscores the urgent aim of improving interoperability across repositories and breaking down silos that prevent information-sharing between items with complementary or even identical provenance that became physically dispersed over time. The *Guidelines* appear at a time when the demand for data-driven, metrics-based dashboards to communicate basic facts and figures about the size and spread of the archival profession has never been greater. The open access re-release of ArchiveGrid in 2012 for the first time allowed archivists and systems developers to analyze the true extent of EAD element usage, via an aggregated nationwide sample of then-124,000 EAD-encoded finding aids [14]. After years of planning, including a persona-based needs assessment [15] and numerous smaller-scale surveys conducted by enterprising research teams with diverse aims and motivations, A*CENSUS II was launched in 2020 to gather comprehensive information about archival worker demographics, educational background, salaries, and perspectives on key issues. In the fifteen years since the first A*CENSUS was fielded, the uptake of even very basic computational methods in archival repositories has shifted the landscape of discovery to an almost wholly digital environment, making the A*CENSUS II a rare opportunity for longitudinal comparisons of repository level-metrics and collections growth.

C. A Tool for Communicating Archival Metrics

The Archival Metrics Project, a preceding five-year effort, developed five questionnaire-based toolkits including one on online finding aids [16, p. 588]. The instruments there support a user studies approach to the evaluation and refinement of archival services and collection access [17]. Decades of archival surveys have investigated not only collection metrics (attributes, patterns, counts, and levels within collections) but also measures of multilingual and multimedia access, the impact of advocacy campaigns and strategic partnerships, and the evolution or versioning of archival policies and standards [18]. A newly established Dataverse, maintained by a standing archival committee on research, data and assessment (CORDA), is making multiple underlying datasets accessible for widespread, innovative reuse [19]. Archival description in the form of finding aids – which capture such elements of a collection as its named identity, scope and content, access conditions, acquisition and appraisal, and the existence of related materials – continues to be interrogated from a number of professional viewpoints, with data visualization replacements [20], alternative approaches [21], and decolonizing prototypes [22] all ready to meet defined challenges in facilitating complex queries of archival records at scale.

III. Practicing Computational Manipulations and Authentic Visualizations

A. Development of IRP2

To ascertain the provenance of both prospective new acquisitions and legacy collections, archivists gather collection-

level and item-level information from subject-specific resources. Museums and archives alike have increasingly recognized the value of institutionalizing internal provenance research programs so as to proactively document the entirety of their collections. Shared resources are an important force multiplier and facilitator of provenance research, given the scale and volume of collections needing such research and its minimization, or even neglect, across professional domains prior to the 1998 Washington Conference on Holocaust-Era Assets, which did promulgate the Washington Conference Principles on Nazi-Confiscated Art amongst 44 countries. Both professional practice communities invested in shared resources development shortly after adoption of the Washington Principles: the American Association of Museums (AAM) launched the Nazi-Era Provenance Internet Portal (NEPIP) in 2003, and under the direction of Michael Kurtz, the National Archives and Records Administration (NARA) launched the International Research Portal for Records Related to Nazi-Era Cultural Property (Portal) in 2011 [23, 24]. The Portal functions as an informational website, listing the resource pages that 18 participating institutions created to enhance public access to their Nazi-era property archives [25]. In early 2015, a research team at the University of Maryland initiated the Portal-enhancement project IRP2 to address the metadata aggregation challenges – indeed, examined since in NAFAN [8] but here more targeted – especially around provenance of the information, Belgian-language searchability, and linked name authorities, that "heirs and families of Holocaust victims searching for lost property, lawyers, investigators, provenance research experts, archivists, museum specialists, librarians, and interested members of the general public" [24, p. 168-9] had reported over the last decade. Milosch, founder of the Smithsonian Provenance Research Initiative (SPRI, est. 2009), echoes Anderson's observations on the fragmentation of art history research in noting the importance of "bringing students into the learning and producing process" [24, p. 166]. Kurtz did exactly that in his role as IRP2 project director from 2015 to 2017, when the resource was responsibly transferred to the European Holocaust Research Infrastructure: EHRI for their ongoing hosting [26].

- 1. Professional Reflections on IRP2 Design, Implementation, Maintenance and Uses for Research

The software developer for IRP2 recalls the IRP2 project director's strong interest to "track looted art, to provide access to art records to the Jewish community, and to make sense of stuff just in such a mess" (GJ, personal communication, April 19, 2023). In particular, it was the international heterogeneity of the Nazi-era looted property universe that had sharpened in the latter's focus since the publication of his book *America and the Return of Nazi Contraband: The Recovery of Europe's Cultural Treasures* (Cambridge University Press, 2006) and the film it partially inspired: *The Monuments Men* (2014) – both of which illustrate the American and Allied investments into cataloging and conducting provenance research post-WWII. For instance, the November 3, 2023 opening of the final permanent exhibit hall of the National WWII Museum in New Orleans, La., which is named Liberation Pavilion, reconstructs an Austrian salt mine housing painting and sculpture masterpiece reproductions and immerses visitors in the *information discovery* processes undertaken then and now by provenance researchers in crisis response settings [27]. The heterogeneity of surviving source data alone posed an "exciting challenge with a fascinating problem to solve, (one) very much in the public interest" (GJ, April 19, 2023) to an experienced software developer nonetheless new to the art provenance world inhabited by museum professionals. Leaders from the SPRI, Art Tracks [28], and the Getty Provenance Index® among others proved willing (to innovate upon traditional ways such as they were) and generous discussants: "we had great meetings about the art provenance issues of the day and whatnot, especially about linking across collections. It was a particular problem that was well-suited from a technical standpoint for the semantic web and linked data RDF: ways of aggregating metadata and making more sense of it" (GJ, April 19, 2023).

Kurtz took the initiative to develop the IRP2 following a dormant period during the 2000s marked by few, or somewhat low-profile, restitutions and growing recognition of a closing window of opportunity for them to occur as people age. Professionally Kurtz had witnessed contract provenance researchers fly across the world to "visit archives to page through in-person to find stuff, and not necessarily find what they want, given the varying levels of cataloging done at different places or not online at all" and the items known only to their staff caretakers (GJ, April 19, 2023). The two workshops hosted by Kurtz's team produced consensus around the ability of IRP2 to reduce the status quo effects of digital erasure and subsequent non-use, and "re-tilt" the research landscape more equitably. By 2017, the "special responsibility" Kurtz showed toward the material was shared by dozens of international collaborators who, even more, revised and reshaped their patron reference activities to be less centered on their particular institution and more cognizant of an interconnected ecosystem of archives worldwide. Ultimately IRP2 is a too-rare exemplar of the transformative problem-solving so needed in many other professional domains too: "Michael was just blown away by what we were able to do!" (see below section), continuing: "I really appreciate that, because to me it was a straightforward-enough technical project that I could sink my teeth into and deliver something, though all I saw at first were the inadequacies of it... But Michael said 'No, this changes everything. Now we have this researching portal that searches for people.' That was a really interesting perspective for me to hear. *And* there's so much to still do." (GJ, April 19, 2023, emphasis added). By drawing attention to the slow pace of progress on property restitution (see [29] on America's role and [30, p. 183] on transitional justice and the mere "twenty-two works" restituted from American museums to victims or heirs between 1998 and July 2006) while addressing the problems from a novel information science perspective, Kurtz lodged IRP2 into the professional settings where it could maximally impact the lives and legacies of millions of individuals persecuted by the Nazis – work that EHRI has been doing since 2010.

Given their "interests aligned," staff from the EHRI visited IRP2 and its partners including the U.S. Holocaust Memorial Museum (USHMM) beginning in 2015, and engaged in discussions to assume the hosting of IRP2: the late Conny Kristel agreed to that on behalf of EHRI (MB, personal communication, March 27, 2023). The development was also supported by an EHRI Advisor and the director of research for the Claims Conference, who has presented comprehensively on several databases making measurable progress: "only in recent years has the opening of archives combined with existing databases and other projects made this possible" [31]. Not only had EHRI "inquired particularly about IRP2 and was always very keen that there should be further development on it" but it happened to meet an internal need of the organization as well: "IRP2 is something that we have never been able to do, just because we haven't had the resources. It's a clear indication that there is definitely interest in the resource" (RS, personal communication, March 27, 2023). The below section further details the team-building and assemblage of skills that Kurtz carried out in his leadership capacity. To expand for a moment on their internal efforts, EHRI researchers had investigated two potential models for searching collections: a federated search which queries many source databases in real-time, and the model implemented in the EHRI Portal where source data are harvested centrally with search results returned from a local repository. Therefore IRP2, in the former way, complements EHRI's offerings to meet a wider range of user needs: "So it's kind of interesting to us on that level. This all relates to how easy it's been to keep it running … The way it does this federated search is to either use an API that is provided officially by the institution, or to scrape their website. … I have done some maintenance on it to try and keep this federated search mechanism working" (MB, March 27, 2023). Based on user behaviors and limited interest in user accounts, EHRI maintainers removed the personalization features and have thus been able to invest into the maintenance of its underlying code, usability for mobile, and compatibility across browsers (that is, five of the six development phases outlined in [24] remain active today). EHRI makes a series of Digital Tools Guides along with code and related resources available on GitHub to facilitate community enhancement, while broadening use, of varied collection data.

EHRI meets a broad range of very precise research needs related to the legacy of WWII by making information available not just about looted art, which is the intention of IRP2, but about archival collections defined and collected based on provenance. Archival collection records in its Portal and in IRP2 do overlap, as both incorporate the linked data resources published by the Getty Provenance Index® to describe places and names. "One of the things we try and do in EHRI is to connect collections based on their provenance. (Aggregation) produces a situation for the researcher when they don't necessarily know if they need to travel (internationally) to review an original collection or if you can just go to the USHMM. So we are trying to connect the metadata of those collections to make the *provenance* of this stuff more clear to researchers. That obviously pertains to the researchers doing art-based research too" (MB, March 27, 2023). A closer look at the IRP2 partner data shows that IRP2 facilitates deep search within the "narrow" category of looted art, and presents multiple levels of granularity of information as search results. "It's very likely that EHRI Portal just has general descriptions of archival collections, pertaining to the Holocaust, held by these institutions. But that query in IRP2 is going to be at much more depth, though much narrower: just [sic] pertaining to information about looted art. … I would imagine that people who do provenance research on looted art would also use the EHRI Portal: it would be useful to pinpoint archival collections that might be of relevance to their work" (RS, March 27, 2023). The EHRI staff are acknowledging the limits of finding aids by its dual offering of the EHRI Portal, which *describes* a collection in about 600 words, and IRP2, which points a user directly to an external database with thousands of records – and they can address those limits by connecting their two resources in future work. Finally they maintain an ongoing interest in raising awareness of IRP2, it having been cited and recognized far too inconsistently: "We don't know what researchers do with the results. Now if it was a digitized collection, it is possible that in the publications they would cite EHRI. But because it's a meta-catalog, the chances that anyone cites the EHRI Portal are very very low, because what happens is they find archival collections via the Portal, but then of course go to the archive to read these things, and then cite the archive and not the Portal. So there is a bit of a problem: we've got a bit of a booking.com problem. That's where people use booking.com to find the hotel they want to go to, but then they book directly with the hotel because it's cheaper. People use the EHRI Portal to find the sources they need for their research, but we actually get very little recognition or acknowledgement afterwards" (RS, March 27, 2023). IRP2's success should lead archivist researchers to their continuing responsibility to articulate to patrons the value of archival labor, so that the seamless services and information harmonized from archival collections are no further invisibilized along the way.

The IRP2 stands as a tremendous multidisciplinary resource facilitating important progress on information-gathering between heirs, belongings, property, and interested parties. It makes connections between otherwise (geographically and functionally) separated pockets of information and facilitates deeply meaningful resolution and reconciliation of such information. Students led major developments of the IRP2 including its federated search capacity, multilingual support, controlled vocabulary-supplied term autosuggestion, and website usability globally [24, 32]. Finally the IRP2 is a strong example of sustained collaboration across educator and practitioner communities, where students play a central role in productive resource development and gain valuable professional experience during their degree.

- 2. Experiential Research Reflections by Then-Student Contributor

As a graduate student in the dual-master's History and Library Science (HiLS) program at the University of Maryland, the second author volunteered to contribute to the development of the International Research Portal for Records Related to Nazi Era Cultural Property (IRP2). The

project was one of many hosted by the College of Information Studies then-Digital Curation Innovation Center (DCIC), a hub of interdisciplinary projects that gave students the opportunity to address archival and information management challenges. From 2015 to 2017, Wachtel collaborated with seven fellow graduate students from the Master of Library and Information Science (MLIS), Master of Information Management (MIM), Master of Human-Computer Interaction (HCIM), and Information Studies doctoral programs. This interdisciplinary nature of the team was vital to the project's success. All brought a variety of skills such as software development, subject knowledge of the Holocaust, German language ability, graphic design, blogging, and coding. Dr. Michael Kurtz's goal for all students involved in this project was to learn valuable technical and collaboration skills. Individual students volunteered for a variety of reasons: Wachtel and her colleague Torra Hausmann sought to contribute their historical subject knowledge and German language skills, colleague Melissa Wertheimer's goal was to help add musical instruments and scores to the types of stolen property vocabulary, and broadly they sought to learn how to enhance access to records of the Holocaust.

One of the most valuable skills they learned as a team was the value of collaboration capitalizing on the individual skills of each student. In addition to regular team meetings, they utilized Asana as a project management tool to spell out each individual's project tasks and progress. In a typical collaboration, Wachtel would work with other MLIS students to suggest a search capability and another student skilled in coding would implement the programmatic feature. For example, in Asana, Wachtel pointed out the pitfalls of utilizing ASCII instead of UTF characters in a multilingual database, which resulted in back-end updates to be able to search using diacritics. In addition, Hausmann and Wachtel suggested user-friendly displays based on their knowledge of archival databases (Fig. 1, Fig. 2). MIM students and their staff software architect Gregory Jansen then translated the metadata for proper display in the final portal design (Fig. 3).

Challenges included the scope of the project across 17 international institutions, with each collections database demonstrating different search capabilities. In a typical collaboration, Wachtel would suggest a search scenario such as "third-generation survivor seeking lost painting." Another student would suggest initial queries and refinements such as "painting" and "Location: Berlin" and MLIS students repeated that search across multiple institutions' collections databases. Team members recorded granular search query metadata across the 17 institutions represented in the portal, including the collection name, a URL of the best search form, query syntax, default operators (such as Boolean operators in German or English), contributing institutions, locations, origin country of the records, language of the

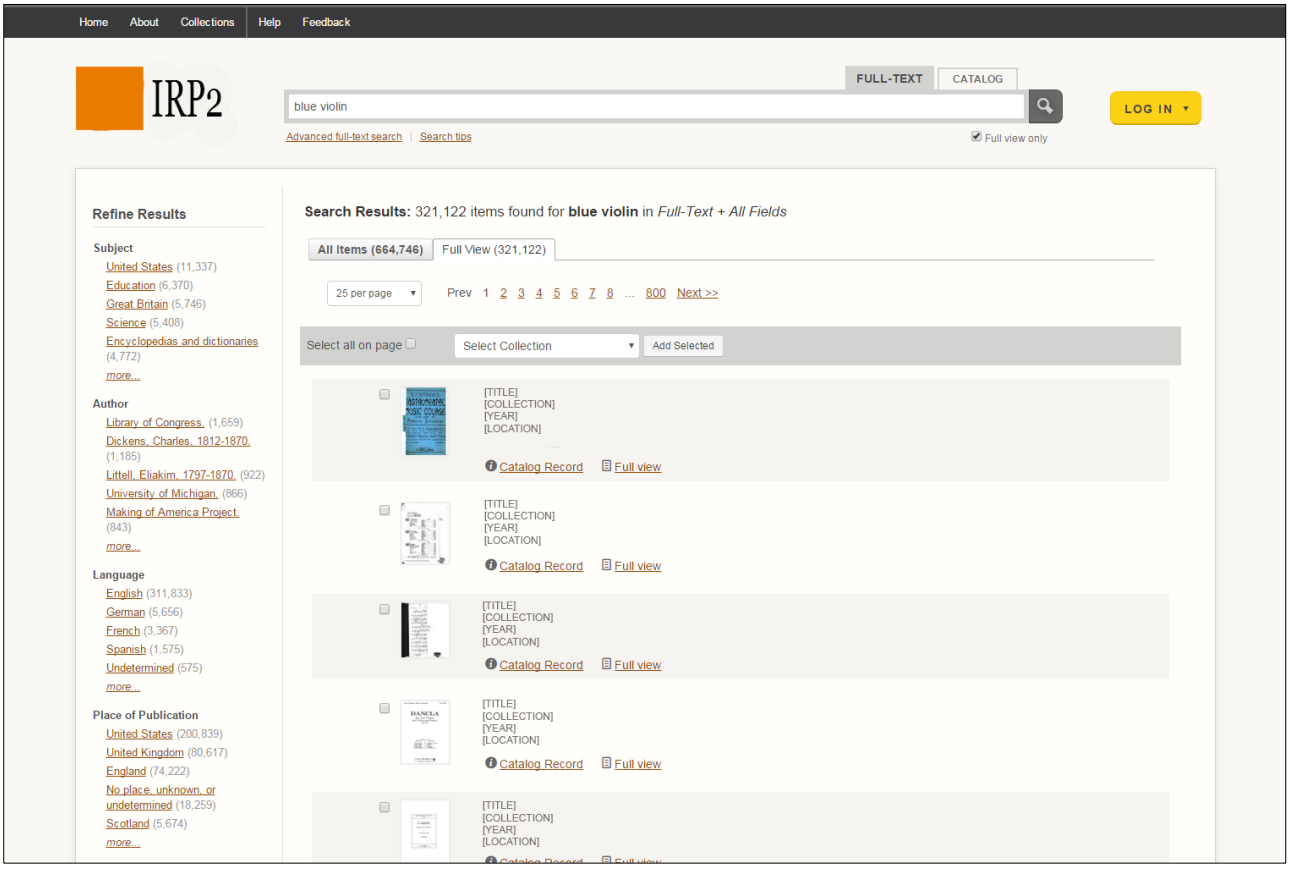

Fig. 2. Mockup of proposed federated search results page layout created by student Torra Hausmann, 2016.

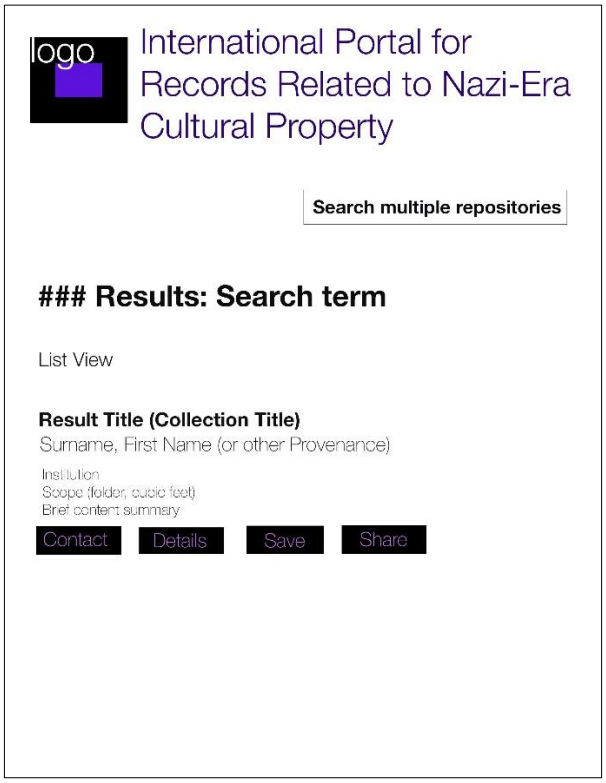

Fig. 1. Mockup of proposed federated search results page layout created by student Jennifer Wachtel, 2016.

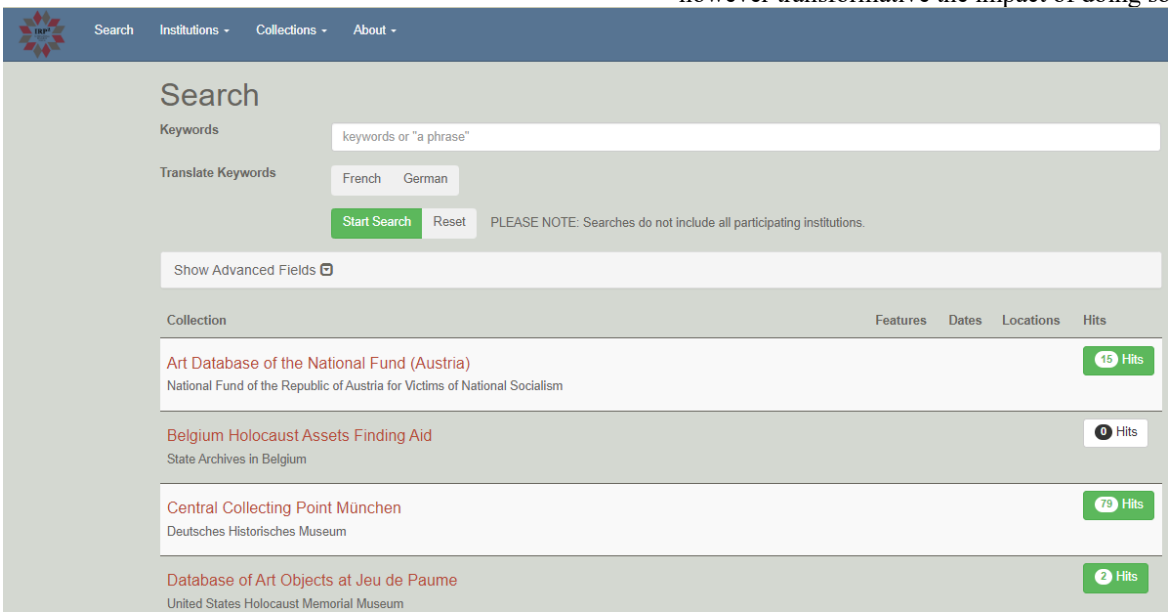

Fig. 3. IRP2 federated search results page, 2023. Courtesy of Jennifer Wachtel.

descriptions or index, records creators, a link to the finding aid, and the kinds of records returned with each search query. Along the way, Wachtel learned valuable lessons about user information needs by testing various user scenarios across multiple public-facing collections databases. Those data were essential to the refinement of search features to improve the portal's usability, especially federated search capacity and controlled vocabularies.

Future archival educators might continue its model of improving access to records through pooled expertise. Dr. Michael Kurtz encouraged and supported each student in their individual goals and coordinated tasks according to their interests and talents. Each member of the team relied on the subject knowledge or coding skills of other team members in order to create the final portal. This project also demonstrated the impact of an environment designed for interdisciplinary research; the development of the portal was achieved through the Center's sponsorship of projects integrating archival research data and technology. Not only does the IRP2 portal potentially profoundly influence the field of provenance research, especially for records related to the Holocaust, but the project offers a model educational approach for practical interdisciplinary collaboration.

B. *CAS Pedagogy with Nuclear Information*

We situate our second pedagogical methods discussion within the core concept of transformative processing that was introduced above, which emerges from an intentional curricular framework. In its design that framework sought to take account of archival work practices today and in so doing embrace a collections as data approach when teaching about digital data formats, ideally prior to directly handling a collection [33]. The framework elevates the applicable lessons from reading Archival Studies beyond the act of processing a collection, however transformative the impact of doing so assuredly is – so that scaling, replicating such efforts become not an afterthought but a retention-conscious mission. Archival Studies graduates emerge with the social and technical abilities to manage hybrid, multimedia collections in a diverse range of settings receptive to the archival enterprise. In addition to its focus on archival practice and inclusion, the curricular framework was infused early on with experiential learning opportunities for audiovisual preservation skills capacity-building among students and graduates [34]. Ongoing integration of active digitization experiences in the Archival Studies emphasis and related courses aids students in their pursuits as new information professionals [35, 36]. Data for this research section were primarily gathered during the Fall 2019 semester, during the second offering of the course IS_LT 9492 Data & Records

Management that is part of the Archival Studies emphasis in a master of library and information science (MLIS) degree program at a North American iSchool.

- 1. Scalable Discovery of Data Patterns

As part of a fieldwide initiative to support archival education and research, a collaboration was established between the faculty instructor and the head archivist at an information agency with a large collection of both digital and analog material. The scalability interests on the part of the archivist become apparent when we articulate that the agency's collection comprises about 500,000 documents (60 million pages), two million photographs, 20,000 film cans, and a subset of other media items including drawings, magnetic tapes, and microfilm. The archivist, this paper's third author, has supervised a remarkably efficient digitization effort resulting in about 20% of the documents digitized and the application of optical character recognition (OCR) and data scraping measures on millions of those pages. Such activities occur as part of a larger machine learning approach to carrying out arrangement and description of the collection [37, 38], first by creating training sets from the metadata extracted from the scanned documents. Complementing the documents themselves is a database of catalog records, for a portion of the documents – each a technical or summary report of nuclear science research. The catalog system is a currently functioning user portal for the agency. The issue of professional interest, though, is that the system uses an in-house version of the standard called COSATI [39] and is exploring the feasibility of continuing to do so or converting the catalog to another standard like MARC [40] or Dublin Core.

In supervising a larger machine learning approach to arrangement and description of the collection, the archivist sought outside perspectives on the content of the catalog records. They generously shared about 26,000 catalog entries with the faculty instructor, using the export tool from the system in place to generate a file in RTF format. They also "cleaned" an initial section of the entries, adding basic formatting to indicate the start and stop of an entry, and to highlight key elements such as its item number, abstract, descriptors, and various dates – all of which we would explore in greater depth. The instructor spent time studying the file and its entries, determining an appropriate amount of material to provide to students who had self-assembled into groups of three. The instructor settled on providing an extract of 125 pages to each of the seven groups (provided in three formats: plain text or TXT, rich text or RTF, and portable document format or PDF). As will be detailed later, that provided students with a couple hundred entries per group, and seemed to strike an agreeable balance between asking students to develop skills and practice using some data manipulation techniques – however basic or advanced [41, 42, 43] – while still being of manageable quantity. The total pages distributed to students represents just 8% of the catalog portion the agency provided us, leaving the instructor with great ability to again draw upon the resource in a later iteration of the course with a different composition of students.

- 2. Learning Objectives

In addition to the dataset described above, students received a data dictionary, presented in the form of an Excel spreadsheet with the full names of each metadata abbreviation found in the catalog records. A few background literature sources and presentation slides about the archival agency and its current collection development efforts enhanced students' grasp of the dataset in context. Three learning objectives structured the students' academic work: (1) Explain the anatomy of a record and the set. What do the records describe? What (perhaps very specific) user needs do they meet? (2) Articulate key content patterns. (3) Summarize the computational manipulations and visualizations you carried out to observe the above patterns. The objectives are informed practically by our dataset attributes and theoretically by the computational archival science (CAS) education framework [44] guiding the collaboration.

- 3. Preliminary Discoveries

In their guest lecture, delivered about one-third into the semester, and another third away from the due date of the assignment, the archivist brought students into the perspective of working to decipher the collection's scope over a period of years. Although their particular work with the collection began more than two decades after 1992, the date of the last U.S. nuclear test, their archival decipherment work is a necessary contribution to the agency's ongoing commitment to nuclear science as stated in the Nuclear Posture Review, then most recently unclassified in 2018 [45]. While the collection is extraordinarily heterogeneous and big, the same may be said of the born-digital dataset of catalog records we were working with. Within each record are one or more descriptors, a field largely supplied by the author of the technical report itself. The previous collection stewards keyed in each descriptor "verbatim" (either from the cover sheet or the text) including any and all instances of misspelling in the original, and cases of faded ink on the paper copy. Only with their hire as the collection's first archivist did any ascertainment of ontologies or controlled vocabularies present in the catalog begin. Apart from descriptor, the data dictionary enumerates about forty other metadata fields with structured or unstructured "tags" in use across the dataset. Certain fields were noted as being of particular interest for the archivist's current collection analysis goals, though the lack of information, or paradata, surrounding the cataloging decisions made by those who handled the catalog over time was clearly noted as well. Students were encouraged to think like a novice user and formulate "computational models to be carried out" based on characteristics as visible as the presence or absence of fields, and discrepancies and duplication of records. Such insights could shed light on provenance traces, or the various layers of processing that had occurred in the decades before our encountering the dataset in 2019.

Students had received their group assignments a week before the archivist's lecture, which gave the students space to prepare and ask specific and informed questions about the dataset. One student was initially concerned about a lack of

familiarity with the subject coverage of the material, and asked about any future release of unclassified material; hearing the archivist talk about their research publication plan was edifying on a number of levels given the students' own planned experiences with research as emerging professionals. Another question addressed the classification statuses of both our dataset and the actual associated reports, and the archivist's answer there helped clarify that while both in our case have been pre-selected at the unclassified "Distribution A. Approved for public release; distribution unlimited" level, much of the non-exported collection portion is classified and in that way occupies primacy in the archivist's day-to-day work. Therefore the contributions our students would be making – pattern-identification and visualizations even if only based on a couple of records (for those students intimidated by working at scale) – could potentially point out hidden or esoteric item characteristics that could apply or even generalize to many other items that could not be shared with us.

A sample of the visualizations generated by the students demonstrates the mutual benefit of this collaboration to the archivist and to all course participants, the positive reception to the data modeling assignment as a component of CAS education, and strategies we might further refine in further course iterations for productive development of problem-solving skills with digital datasets. Fig. 4 shows a geographical heat map of the kind promoted by the ArchivesZ researcher-developers. Students noted that unlike author names or various numerical identifiers, corporation names were visualization-ready because their list of names (facilitated by OpenRefine and task delegation among group members) could then be located on a map (called a choropleth) and then color-shaded to illustrate how scattered or isolated were the places of origin for our collection of reports. Fig. 5 shows the desired investigation into how many descriptors were present in a given catalog record: we see a mean of 12, a median of 11, and also a mode of 12 (appearing 19 times in their sample); all elements were obtained using the count function in Excel. Based on their analytic discovery, students concluded "that the majority of authors included up to 5 (or 0) descriptors. However, authors commonly included up to 20 descriptors." In general, students readily acknowledged how and in what ways their visualizations were limited by nature of the randomness in working with one portion of a much larger dataset (e.g. pie charts would compound an illusion of completeness and thus were avoided), as well as their consideration of the specific information needs that could be met through their work of observing, manipulating, and visualizing the data.

- 4. Reflection, and Incorporation of Student Findings by the Archivist

The sharing of student findings about their collection work with the professional archivist informed the archivist's presentations of the collection to potential researchers. Specifically, the students surfaced depth and quantity patterns that revealed greater amounts of certain material than were thought to be there, prompting revised catalog metadata subject fields. The visualizations provide greater context into the black box of unknowns associated with the

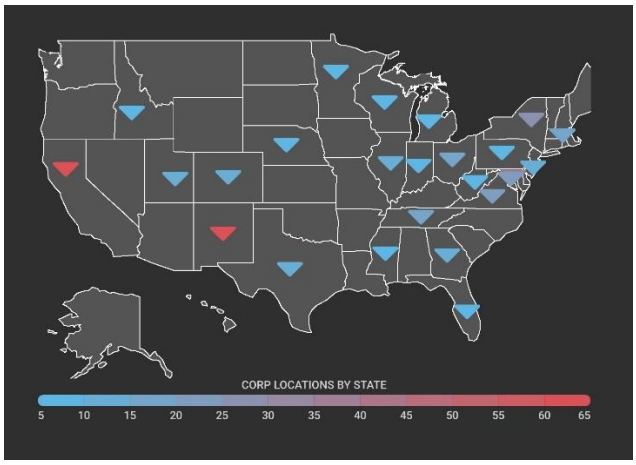

Fig. 4. Choropleth map of the data sample, 2019. Image courtesy Jane Doe.

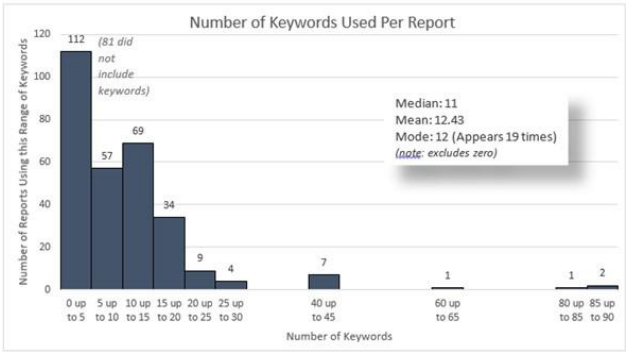

Fig. 5. Bar graph showing descriptor counts within records, 2019. Image courtesy Jane Doe.

collection. The students' work established ground truth from which two different machine learning pathways were developed by the third author: an annotation study to create a 'gold set' of data and a data dictionary that identified all acronyms.

The students' work illustrated that a high number of keywords were used as descriptors. That insight was twofold: it signaled to the archivist that there is specific scientific content required for users and that it would be critical for subject matter experts to be involved during the data curation machine learning exercises. Ultimately the students' analyses told the archivist that the current system was no longer usable. Visualizations finally provided the proof and support for data-driven decision making – later estimated as three months of work time – to export the intact metadata in XML format and operationalize a new more accessible system. Precise researcher queries of the information were not affected, as the new user interface accesses cleaned metadata and OCR-digitized documents. The visualized data were the beginning of the entirety of the collection being analyzed semantically.

IV. Discussion of Outcomes

Visualization has been explored here as a data practice, a creative response to archival metrics, and a tool for communicating attributes, patterns, counts, and levels within collections. Most practically connected to the activities of

archival arrangement as a step toward description and even descriptive aggregation, we contend that visualization-based data modeling activities have the potential to expose students to a shared transformative processing experience. Depending on the dataset selected, which in our second case was a born-digital word-processed file of catalog records, the design and execution of a data modeling assignment in the archival classroom may allow groups of students to become familiar with distinct aspects of the data collection. For example in a course combining online and in-person students, such students may be able to focus on visualizing a collection's metadata and materiality respectively. A limitation of this research is a relative lack of attention to exploring the materiality of arrangement work, as prior research has brought forth [46]. In the context of a blended classroom modality where some but not all students in the course gathered physically to attend class, most of the data modeling steps, micro-discoveries, and progressions were carried out virtually on primarily textual and even numerical "records," with minimal visual elements to start with. Such a dataset lent itself to visualizing patterns in metadata (including intangible information such as subjects) but few physical characteristics. That knowledge remains with the archivist partner whose knowledge of the collection's physical contours, when combined with the students' metadata pictures, will inform future presentations of the material to patrons. Still the resulting visualizations and accompanying narratives were produced in a highly coordinated manner and even compared and combined with others' in making a summative deliverable, as related studies of archival arrangement practices and data modeling have called for [47]. A key problem of arrangement is that it frequently becomes a transactional activity for what becomes the definitive representation of a collection for public access, belying the proper attention archivists might give to trialing one or more visualization strategies to meet distinct user needs, or across the archival lifecycle [48]. Visualization exercises permit archivists to test and prototype more or less optimal collection representations to broaden interest in the material, and students with a powerful role in the same.

V. Ongoing Directions

While anticipated in advance, the born-digital, non-visual nature of both datasets allowed students to focus on activities such as proper names, subject classification, date frequency, and geographic spread at the possible expense of incorporating extrinsic materiality into a processing plan. Perhaps relatedly, neither of the goals of our data processes respectively were appraisal, but rather to allow student archivists to explore digital representation(s) of a stable collection as currently presented to prospective, subject-expert users with precise information needs. Future work with other kinds of datasets may be better equipped to direct differently focused kinds of exploration. In considering the goals and context of the data modeling activity depicted, the student archivists gained hands-on experience with an active dataset exported from a live system, one which a professional archivist is tasked with managing. Though data modeling remains at best an emerging area of interest among the current Archival Studies student body, completion of such authentic analysis serves as proof, to themselves as well as those who will evaluate their work, of one's flexibility in learning technical skills and commitment to peer mentorship and teamwork. Those kinds of aptitudes can transfer across professional domains and persist within an archival managerial career.

Acknowledgments

The authors gratefully thank all participants in the two courses for their support, and dedicate this work to our late colleague Michael Kurtz. An early version of Section III.A.2. was presented at the Mid-Atlantic Regional Archives Conference: MARAC 2017 in Buffalo, NY [23], and preliminary work for Section III.B.1. was presented at the Archival Education & Research Institute: AERI 2019 in Liverpool, United Kingdom [38].

References

[1] O. W. Holmes, "Archival arrangement: Five different operations at five different levels," American Archivist 27(1): 21-42, Jan. 1964. doi:10.17723/aarc.27.1.l7218571l7617w15

[2] M. A. Puente, facilitator, "Reaching in to reach out: Examining the state of inclusivity across libraries, archives, and museums," Presentation at ARCHIVES 2017. Portland, Oreg., 27 July 2017. https://sched.co/AG68

[3] T. Vancisin, L. Clarke, M. Orr, and U. Hinrichs, "Provenance visualization: Tracing people, processes, and practices through a data-driven approach to provenance," Digital Scholarship in the Humanities 38 (3): 1322-1339, 2023. doi:10.1093/llc/fqad020

[4] E. Ragan, A. Endert, J. Sanyal, and J. Chen, "Characterizing provenance in visualization and data analysis: An organizational framework of provenance types and purposes," in Proceedings of IEEE VIS: Visualization & Visual Analytics, VAST Conference: Visual Analytics Science and Technology (pp. 31-40), Chicago, Ill., 27 October 2015. doi:10.1109/TVCG.2015.2467551

[5] D. J. Wrisley, "Pre-visualization," in Proceedings of IEEE VIS: Visualization & Visual Analytics, 3rd VIS4DH Workshop, Berlin, Germany, 21 October 2018. https://vis4dh.dbvis.de/2018/

[6] S. A. Matei and L. Hunter, "Data storytelling is not storytelling with data: A framework for storytelling in science communication and data journalism," The Information Society 37 (5): 312-322, 2021. doi:10.1080/01972243.2021.1951415

[7] L. Saffran, S. Hu, A. Hinnant, L. D. Scherer, and S. C. Nagel, "Constructing and influencing perceived authenticity in science communication: Experimenting with narrative," PLoS ONE 15 (1): e0226711, 2020. doi:10.1371/journal.pone.0226711

[8] C. S. Weber et al., Summary of research: Findings from the Building a National Finding Aid Network Project. Dublin, Ohio: OCLC Research, 2023. doi:10.25333/7a4c-0r03

[9] M. Whitelaw, "Towards generous interfaces for archival collections," Comma: International Journal on Archives 2012 (2): 123-132, 2012. doi:10.3828/comma.2012.2.13

[10] J. Joque, A. Zoss, and A. Rutkowski, Visualizing the Future symposia: An IMLS funded National Forum on data visualization in libraries, 2018. https://visualizingthefuture.github.io/

[11] S. Braun, "Critically engaging with data visualization through an information literacy framework," Digital Humanities Quarterly 12(4), 2018. http://www.digitalhumanities.org/dhq/vol/12/4/000402/000402.html

[12] M. O'Hara Conway and E. R. Novak Gustainis, "Counting in a common language: Standardized holdings counts and measures for archival and special collections repositories," Archival Outlook: 13, 16, Jan.-Feb. 2020. https://www2.archivists.org/archival-outlook/back-issues/2020

[13] J. Golbeck, "Visualizing archival collections," NEH Digital Humanities Start-Up Grant White Paper, 2010. https://securegrants.neh.gov/publicquery/main.aspx?f=1&gn=HD-50505-08

[14] M. Bron, M. Proffitt, and B. Washburn, "Thresholds for discovery: EAD Tag analysis in ArchiveGrid, and implications for discovery systems," Code4Lib Journal 22, 2013: https://journal.code4lib.org/articles/8956

[15] M. Light, chair, "Recommendation to establish a Committee on Research, Data, and Assessment." Prepared for the November 2-3, 2018, Society of American Archivists Council Meeting, Chicago, Ill., 2018. https://www2.archivists.org/groups/saa-council/november-2-3-2018-council-meeting-agenda

[16] W. Duff et al., "The development, testing, and evaluation of the Archival Metrics toolkits," The American Archivist 73(2): 569-599, 2010. doi:10.17723/aarc.73.2.00101k28200838k4

[17] B. Craig, "Perimeters with fences? Or thresholds with doors? Two views of a border," American Archivist 66(1): 96-101, Spring/Summer 2003. doi:10.17723/aarc.66.1.a6773715p68m2068

[18] J. Furner, "'Records in context' in context: A brief history of data modeling for archival description," in F. Foscarini, H. MacNeil, B. Mak, & G. Oliver (Eds.), Engaging with records and archives: Histories and theories (pp. 41-62). London: Facet, 2016. doi:10.29085/9781783301607.004

[19] P. Conway, "Archival research: Past & future," in Proceedings of the 2020 SAA Research Forum, August 5, 2020. https://www2.archivists.org/am2020/research-forum-2020/agenda

[20] A. Bahde, "Conceptual data visualization in archival finding aids: Preliminary user responses," portal: Libraries and the Academy 17(3): 485-506, 2017. doi:10.1353/pla.2017.0031

[21] J. G. Daines III and C. L. Nimer, "Re-imagining archival display: Creating user-friendly finding aids," Journal of Archival Organization 9(1): 4-31, 2011. doi:10.1080/15332748.2011.574019

[22] L. Haberstock, "Participatory description: Decolonizing descriptive methodologies in archives," Archival Science 20(2): 125-138, 2020. doi:10.1007/s10502-019-09328-6

[23] M. Kurtz, N. Yeide, J. L. Wachtel, and G. Jansen, "Enhancing access to Holocaust records: Developing a prototype International Research Portal," Presentation at the Mid-Atlantic Regional Archives Conference: MARAC. Buffalo, N.Y., 27 October 2017. doi:10.13016/M2000019F

[24] J. Milosch, M. J. Kurtz, G. J. Jansen, A. Hull, and R. Marciano, "Museums, archives, and universities—Structuring future connections with big data," in G. Schiuma and D. Carlucci (Eds.), Big data in the arts and humanities: Theory and practice (pp. 159-172: ch 13). Routledge Taylor and Francis Group, 2018. https://ai-collaboratory.net/cas/cas-publications/

[25] U.S. National Archives and Records Administration, "International Research Portal for Records Related to Nazi-Era Cultural Property," Washington, D.C., 2021. https://www.archives.gov/research/holocaust/international-resources

[26] European Holocaust Research Infrastructure, "EHRI hosts new online resource: The International Research Portal for Records Related to Nazi-Era Cultural Property," news release, 13 June 2017. https://www.ehri-project.eu/ehri-hosts-new-online-resource-international-research-portal-records-related-nazi-era-cultural

[27] C. Karpiak, "Top museum in the country announces new immersive post-war Liberation Pavilion," MediaDecision, 16 June 2023. https://www.mediadecision.com/news/state/louisiana/top-museum-in-the-country-announces-new-immersive-post-war-liberation-pavilion/article_4ea274a8-fd7d-11ed-b6eb-9f337580d907.html

[28] Carnegie Museum of Art, 2016 Digital Provenance Symposium. Pittsburgh, Pa., 14 October 2016. https://www.museumprovenance.org/pages/scholars_day_2016/

[29] L. Foulkes, "The art of atonement: How mandated transparency can help return masterpieces lost during World War II," Boston College International & Comparative Law Review 38: 305-328, 2015.

[30] B. Demarsin, "Let's not talk about Terezín: Restitution of Nazi Era looted art and the tenuousness of public international law," Brooklyn Journal of International Law 37(1): 117-185, 2011. https://brooklynworks.brooklaw.edu/bjil/vol37/iss1/3/

[31] W. A. Fisher, "Twenty years after Washington: An evaluation from the Claims Conference and the WJRO," in Provenance & Research (pp. 17-20). Magdeburg, Germany: German Lost Art Foundation, 2020. ISBN 978-3-95498-651-4. https://art.claimscon.org/resources/resources-reference-list/

[32] L. M. Rogers, "The IRP2 [logo]," Lisa Marie Rogers: User Researcher, 2017. https://lisamarierogers.com/2017/01/04/the-irp2/

[33] V. Ikeshoji-Orlati, "Vanderbilt Library Legacy Data Projects," Collections as Data Facet #13, 2017. https://collectionsasdata.github.io/facet13/

[34] A. Roeschley, S. A. Buchanan, M. Burke, A. Graf, and O. L. Zavalina. "Considering individual and community contexts within information pedagogy, scholarship, and practice," in Proceedings of the Association for Information Science & Technology 57: e283, 2020. doi:10.1002/pra2.283.

[35] M. Villesvik, "PBPF Program Equipment Plans," in Blog of the American Archive of Public Broadcasting (AAPB), a collaboration between WGBH and the Library of Congress, 2018. https://americanarchivepb.wordpress.com/2018/12/21/pbpf-program-equipment-plans/

[36] C. Davis Kaufman, J. Elmborg, R. Fraimow, T. Schneiter, and M. Ulinskas, "Toward a more equitable field: Broadening the landscape with fellowships in audiovisual preservation," Journal of Archival Organization 17(1-2): 19-37, 2020. doi:10.1080/15332748.2020.1769995

[37] J. Stevenson, "A supervised machine learning approach to arrangement and description," in Proceedings of the 2018 SAA Research Forum, Washington, D.C., August 14, 2018. https://www2.archivists.org/am2018/research-forum-2018/agenda

[38] J. Dorey and J. Stevenson, "A weapon of math destruction: Implementing machine learning for archivists in research and practice," Workshop at Archival Education & Research Institute: AERI. Liverpool, United Kingdom, 9 July 2019. https://aeri.website/

[39] M. M. Henderson, "The data element dictionary: Developments in technical reports processing," Government Publications Review 13 (5): 581-590, Sep-Oct. 1986. doi:10.1016/0277-9390(86)90050-6

[40] A. G. Hoshovsky, "COSATI information studies – what results," in Proceedings of the Annual Meeting of the American Society of Information Science (ASIS), San Francisco, Calif., 3 October 1969. https://apps.dtic.mil/sti/citations/AD0695476

[41] G. Wiedeman, "XQuery for archivists: Understanding EAD finding aids as data," Practical Technology for Archives 3, Nov. 2014. https://practicaltechnologyforarchives.wordpress.com/issue3_wiedeman/

[42] J. Rider, "Repurposing archival metadata with the Python CSV Writer," in Programming for Cultural Heritage (PFCH), A Pratt Institute School of Information course, 2016. http://pfch.nyc/repurposing_archival_metadata/index.html

[43] W. Underwood and R. Marciano, "Computational thinking in archival science research and education," in Proceedings of IEEE International Conference on Big Data, 4th CAS Workshop (pp. 3146-3152), Los Angeles, Calif., 11 December 2019. doi:10.1109/BigData47090.2019.9005682

[44] R. Marciano, G. Jansen, and W. Underwood, "Developing a framework to enable collaboration in computational archival science education," in Proceedings of the 2019 SAA Research Forum, Austin, Tex., August 2, 2019. https://www2.archivists.org/am2019/research-forum-2019/agenda#peer

[45] U.S. Department of Defense, "Nuclear Posture Review," Washington, D.C., 2018. https://dod.defense.gov/News/SpecialReports/2018NuclearPostureReview.aspx

[46] C. B. Trace and L. Francisco-Revilla, "The value and complexity of collection arrangement for evidentiary work," Journal of the Association for Information Science and Technology 66 (9): 1857-1882, 2015. doi:10.1002/asi.23295

[47] K. M. Wickett, "Critical data modeling and the basic representation model," Journal of the Association for Information Science and Technology, 2023. doi:10.1002/asi.24745

[48] A. Gilliland and R. Marciano, "Preparing archivists in computational thinking and innovative technologies," Keynote for Technology, Society, Humanities: International Academic Conference on Digital Intelligence Empowering the Modernization of Archival Work. Shandong University, China, 26 October 2023. https://youtu.be/ZdLJHQLbR4k