VRA Bulletin

Volume 46
Issue 2 *Fall/Winter 2019*

Article 4

May 2020

Images as Research Data and the Role of the Information Professional

Jasmine Burns
Cornell University, jeb467@cornell.edu

Images as Research Data and the Role of the Information Professional

Abstract
The goal of this paper is twofold: first to define "research data" in a humanities context through a discussion of the ways in which humanities researchers create and aggregate image collections, and second to address the processes by which academic libraries and information professionals can play an active role in supporting the treatment and perception of images as research data. Much of what the author discusses in this capacity can be applied across the Visual Resources community, regardless of departmental affiliation.

This article has undergone a double-blind peer review process.

Keywords
metadata, advocacy, image management, visual resources, research data

Author Bio & Acknowledgements
Jasmine Burns is the Visual Resources Metadata Librarian, Cornell University Library.

Humanities Data

Recent and rapid advances in technology have drastically changed the ways in which humanities researchers access and engage with materials for their research. There is a great deal of literature on the history of technology and digital materiality that explains the cultural and material shifts that led to this change. Much of this scholarly work highlights the notion that more people than ever carry a high-quality camera in their pocket everywhere they go, and due in part to this technology, we as a society have amassed an enormous volume of images and visual materials that are shared openly online. Almost everything we encounter is mediated by some kind of digital technology.[1] For more literature on this topic, it would be useful to look toward the field of the History of Science and Technology, as well as the wealth of literature on digital culture /new media and digitization in the cultural heritage sector.

The goal of this paper is twofold: first to define "research data" in a humanities context through a discussion of the ways in which humanities researchers create and aggregate image collections, and second to address the processes by which academic libraries and information professionals can play an active role in supporting the treatment and perception of images as research data. Much of what I discuss in this capacity can be applied across the Visual Resources community, regardless of departmental affiliation.

From Google images to iPhone snapshots in museums and archive reading rooms, the contemporary humanities researcher compiles large digital image collections that become essential to their scholarly work. For the purposes of this paper, I define image collections as any group of image files aggregated by a researcher regardless of content (this could include visual, textual, and non-representational materials). The examples in this paper are grounded within the visual arts, but can be widely applied to humanities research generally. Visual arts research data is defined by Garrett & Gramstadt as "data which arises out of, and evidences, research. This can be classified as observational, including: sensor data; experimental; simulation; derived or compiled data for example databases and 3D models; or reference or canonical for example, a collection of smaller datasets gathered together. Examples of visual arts research data may include sketchbooks, log books, sets of images, video recordings, trials, prototypes, ceramic glaze recipes, found objects, and correspondence."[2] Just as a biologist's research may rely on a set of photographs documenting the stages of larval fish development, an architectural historian's work likely depends on a series of plans and sketches. However, the terminology that defines "research data" in the STEM fields does not readily apply to the humanities, and because of the many processes of collecting and interpreting data in the humanities fields, it becomes very difficult to pin down a single definition: data in the humanities can look, sound, and be almost anything. As a result, humanities researchers have a tendency to resist the term, claiming that they do not produce research data and therefore have no need for data management strategies or workflows.

Additionally, the difference in the treatment of research data between STEM and the Humanities is largely cultural: we as a community of researchers and information professionals do not systematically deposit color samples or architectural plans, for example, into an institutional repository to accompany a piece of scholarship in the same way that a STEM researcher would think (or often be required) to deposit their raw data to be openly accessible. This is not to suggest that these materials do not belong in any of these repositories, but rather that the drive to preserve and share research data in this way must start with the researcher and largely depends on *their* understanding of those materials as research data from the outset.

Creating and Aggregating Image Collections

Throughout the research process, scholars become curators of their very own virtual collections. Two common methods for collecting visual materials are to take photographs in situ or to download

[1] As a disclaimer, I am also working under the assumption that there are certain privileges afforded to those working within academia with regards to access to these technologies.

[2] Garrett, Leigh and Gramstadt, Marie-Therese (2012) KAPTUR: exploring the nature of visual arts research data and its effective management. EVA LONDON 2012 Electronic Workshops in Computing, 90.

images from a variety of online sources. These image files can either stand alone and constitute data or they can serve as a surrogate for the data source itself, which is then contextualized through description, i.e. descriptive metadata. An example of the former is the "movies in color" project,[3] where image files are analyzed as grids of color in order to enable new modes of film visualization and analyses of set design in relation to color palettes. An example of the latter would be photos taken in an archive reading room or on site in a museum, etc.

There have been several surveys that address the ways in which researchers aggregate and organize their image collections. These surveys are obviously very limited in their scope and are therefore not comprehensive, but I have chosen to use them here as a guiding framework to open up a discussion of the range of practices and tools available, rather than as a quantitative data point. In ITHAKA's "Supporting the changing research practices of Art Historians,"[4] more than 70 scholars from a variety of institutions were interviewed about their research needs and strategies. One of the largest changes cited in the findings was the use of digital images for teaching and research. Although these personal image collections have become integral to the work of art historians, ITHAKA claims that they "have not yet systematized the processes to organize and manipulate these files, and they do not always have the right tools available to manage their personal research collections." This statement may be an over-exaggeration, as I have worked with art historians who have indeed found tools to successfully manage their image collections, but it draws out an issue that can and should be used to drive more development towards the creation of better and more effective tools.

Within the same interview series, ITHAKA conducted a survey focused on historians.[5] This survey found that because of digital cameras, research trips undertaken by historians have morphed into data collection missions, rather than time spent pouring over documents for immediate analysis in the archive reading room.[6] This is a method that I have seen used by many types of scholars, and is not limited to historians. Both surveys note that most scholars are using basic folder structures while relying on memory to organize their collections, and are placing crucial metadata into the folder and filenames to facilitate searching. According to the art historian survey, "a very small minority" of researchers (based on the limited 70 plus subjects of the study) use any other tools or applications for organization.

A similar survey conducted by the developers of Tropy[7] (a platform to be discussed later in this paper) noted that of the 110 scholars they interviewed, most were generally dissatisfied with these kinds of homegrown organizational systems. Training and development in this area is mostly ad hoc as there has been little done to prepare current and emerging scholars for the transition into using digital tools and the anxieties it produces.[8] While these ad hoc strategies for file management are neither ideal for the researcher's current uses, nor for the future archivist who will process the researcher's papers, it does provide an opportunity for visual resources and metadata professionals to offer solutions to file organization.

Role of the Information Professional

[3] Movies in Color, accessed August 2019, http://web.archive.org/web/20190724150213/https://moviesincolor.com/
[4] Schonfeld, Roger and Matthew Long. "Supporting the Changing Research Practices of Art Historians." Ithaka S+R. Last Modified 30 April 2014. https://doi.org/10.18665/sr.22833.
[5] Rutner, Jennifer and Roger Schonfeld. "Supporting the Changing Research Practices of Historians." Ithaka S+R. Last Modified 7 December 2012. https://doi.org/10.18665/sr.22532.
[6] A major benefit of this shift in research practice has been to allow researchers without the resources to undertake time-intensive trips to undertake archival research. For a more in-depth discussion of this, see Burns, "Information as Capital: The Commodification of Archives and Library Labor."
[7] Mullen, Abby. "Untangling the Mess: Researchers' Photo Practices." Tropy Blog. Accessed July 2019. http://web.archive.org/web/20190730200144/https://tropy.org/blog/untangling-the-mess-researchers-photo-practices/.
[8] Thomas, William G., "Renegotiating the Archive: Scholarly Practice in the Digital Age." Personal Blog, 2016. https://web.archive.org/web/20190401185505/http://railroads.unl.edu/blog/?p=1195

Not only do humanists create and store research data in a range of formats, but many researchers grapple with being underprepared to manage, access, and disseminate these data in a digital environment. Academic libraries and information professionals are therefore ideally situated to inform the best practices for humanities research data management. They are also well-equipped to provide the infrastructure, tools, and training for managing such collections, and to make these data (i.e. image collections) available and useful to a wide range of users.

The main conclusions of the authors of the aforementioned surveys shift the focus onto "the next generation" of scholars and recommend addressing the current lack of training at a graduate level. I would even go so far as to say that information professionals could start with undergraduates and incorporate digital tools and organizational strategies into instruction and /or consultations.[9] A more sustainable method may be to collaborate with faculty to create required assignments that are embedded into the curriculum. This kind of embedded work (where the information professional plays a central role in the classroom throughout the semester) is trending in some institutions in the US. For example, at Indiana University Bloomington, the library provides small information literacy grants to facilitate this kind of faculty-librarian collaboration. In addition, at my own institution we have a librarian embedded into the School of Architecture, Art, and Planning.

Metadata

Information professionals can provide a great deal of support for research data management through discussions of metadata. Metadata is customized for organization, access, and description through the application of different schema, but the overall structures and standards change very little across disciplines. Useful metadata comes from the consistency of application as well as description across a given corpus (be that a collection or a system). Humanities service providers could offer better support to humanities researchers by taking inspiration from the support services already available to the STEM disciplines. For example, Cornell's Research Data Management Service Group[10] has an excellent website with best practices, definitions, and full consultation services (mainly focused around STEM but they do sometimes work with humanities scholars who seek out their services). Although this service exists on campus, and is well positioned to assist with the management and organization of image collections, it is not a logical service point for, say, a printmaking faculty member to manage their students' digital portfolios.

During my own consultations within my campus community, I always reiterate that the key to a well-organized collection of images is consistency. This can be achieved through good documentation and clear definitions, but does not necessarily mean following a complicated schema with lots of rules and standards. This is particularly significant for those people who are managing personal image collections, where the user and the creator of the collection are one in the same. The most important guiding principle in this case is to establish a set of metadata fields and to clearly define their use. I avoid trying to push library-level standards onto casual users.

As an example, when setting and defining fields, "Date" could indicate the date that the source material was created as opposed to the date that the photograph was taken, or "Location" could be either the geographic location of a public work /site or the actual name of the repository or institution where the item is held, etc. These definitions are particularly significant if and when the software or platform being used for organization or access does not have customizable fields and you have to make do with what is available. A significant consideration when creating this master list of fields and definitions is to account for all possible types of data for that particular project or collection, because all sorts of problems will arise later on if new fields need to be added, or the existing data has to be parsed out. We, as information

[9] That is not to say that we are not already engaging with this kind of instruction already, or should struggle to cram a huge amount of information into these sessions, just that these strategies could form the basis of most of our interactions with students.

[10] https://data.research.cornell.edu/

professionals, do not have to push metadata schema and standards onto users who will find little or no use for them outside of a library context. When working with researchers, I have found that the best approach is to meet them at their point of resistance: what do they need and why is their current practice not meeting those needs.

When explaining how to make use of metadata, it is helpful to keep workflows simple and clear. Most researchers find very little use in elaborate schemas as many already struggle with finding sufficient data to populate fields. It can be useful to start with an established structure, but to also allow deviation from the named schema upon which your fields are based. However, one of the major distinctions that I try to highlight is the difference between metadata and research data: metadata being data that represents the image or the file itself, so descriptive and /or technical data, and research data as an interpretation of the content or context of the image. This becomes very messy in the humanities because of the overlap between interpretation and visual analysis. However, with clear field definitions, researchers can easily distinguish the types of data they are collecting and start to form their own taxonomies.

Tools

Changing the perception of "a bunch of images on my desktop" into "this is my research data" is only half the battle. The other half is suggesting and implementing practical solutions for image and metadata management. There are several tools that I have suggested for this kind of personal collection management. As a note, these examples do not include methods for providing wider access to collections, nor for implementing digital projects; however, if used consistently, these tools will help facilitate those goals. This is by no means an exhaustive list of the tools available, and there is not one platform or tool that fits all image management needs. Some projects may even require a combination of tools and techniques. The goal of this section is to give a brief introduction to several popular solutions for managing personal image collections as data.

Tropy is an open-source desktop platform, developed by the Roy Rosenzweig Center for History and New Media,[11] that is specifically designed for scholars conducting archival research. It is an excellent platform for organizing and annotating images, and I would recommend this for many different types of projects. It would be best utilized for people who are starting a new project, as integrating an ongoing project would be laborious, though not impossible. Their online documentation is robust, and their discussion forums are very useful for troubleshooting. The Tropy developers are also highly amenable to community-suggested enhancements /changes /updates for future releases.

Pros:

1. The platform is self-contained on the user's desktop, where they can import individual image files.
2. The metadata fields are customizable. While they are selected from an existing schema, they can be custom-organized according to the researcher's needs.
3. The platform utilizes existing controlled vocabularies.
4. It supports item, photo, and detail-level cataloging.
5. There are multiple mechanisms for sorting /organizing images and data within the platform, including metadata templates, tagging, grouping multiple images into complex objects, etc.
6. It offers basic photo editing functions (crop, rotate, zoom, sharpen).
7. It allows export to Omeka S for the creation of online exhibits.

Cons:

1. It does not fully support the aggregation of external sources, so users can only import image files (JPG, PNG, SVG, GIF, TIFF), and not the embedded or existing metadata that may be attached to those files. That information has to be input manually and there is no batch import function.

[11] https://tropy.org/

2. The data exist as a single Tropy file that currently only exports as a JSON-LD file with all metadata and images. The issue here is that users may not know what to do with a JSON file or how to parse out images and metadata.

Zotero is another open-source platform developed at the Roy Rosenzweig Center. It is designed as a text-based citation management tool, but it can be leveraged to be used with images. Its greatest strength is that it reuses existing metadata from web sources, and would therefore be most useful for someone who needs to create consistent citations and /or links to found online images. Alexander Watkins wrote a handy guide to using Zotero for image management in a 2013 article published in *Art Documentation*.[12] While many of the platforms mentioned in this article are now defunct or have changed significantly, the practical guide for using Zotero to store and manage images is much the same. It is important to keep in mind that Tropy, in many ways, is a response to the needs of personal image management that Zotero was not able to fulfill.

Pros:
1. Metadata fields are structured and defined, but there is still space for descriptive metadata.
2. The tool is partially relational, meaning it can create links between objects.
3. Collections are synced with cloud storage and are retrievable via a web interface.
4. It will automatically generate image citations.
5. There is a series of plugins, or translators, that tell Zotero how to pull information and files from individual databases, websites, or catalogs. For instance, there is an Artstor translator, which allows users to import the file and metadata from the Artstor Digital Library directly into their Zotero library with just one click. These also exist for a number of image databases, and are customizable with a little bit of web scraping and HTML knowledge.[13]

Cons:
1. The metadata field labels are restrictive, i.e. not editable.
2. The metadata fields are not mapped to a known schema for interoperability.
3. The files are stored in the application and not externally, so periodic local backups would be necessary.

Adobe has created a suite of image management tools designed for professional photographers that includes Photoshop, Illustrator, Bridge, and Lightroom. Bridge and Lightroom in particular have the capability to embed metadata into image files, which can be leveraged for managing assets that are part of an existing file structure. Embedded metadata allows users to search for any of this embedded text directly from the desktop search bar (in both Mac and PC). Users should maintain documentation of how they are using the fields (i.e. clearly define their fields), and keep standardized vocabulary lists for consistency.

Pros:
1. This tool uses existing file structures, so users do not have to adapt to a new organization system.
2. It has reusable and customizable embedded metadata templates that can be applied in batches.
3. Bridge can be used to batch rename files.

Cons:
1. The metadata fields follow the IPTC Core schema and are therefore highly restrictive and specific to photography practice.

[12] Watkins, Alexander. "Zotero for Personal Image Management," *Art Documentation* (Fall 2013): 301-313.
[13] The code for which is available here: https://github.com/zotero/translators/blob/master/ARTstor.js

2. It was initially created for professional photographers, making it somewhat challenging to adapt to a wide range of use cases.
3. There could be a significant learning curve when users are first starting out.

A more customized approach would be to use a structured database, such as Microsoft Access or FileMaker Pro. These are highly flexible and customizable tools that are used universally across disciplines. Much of the functionality that I have highlighted in the above examples is available in these options, and it makes it very easy to share, export, and publish data that are structured through these systems.

Pros:
1. These are excellent options for complex data collections that contain multi-layered information.
2. They are fully relational.
3. They are fully customizable (i.e. metadata fields, cataloging interface, exportable data structures).

Cons:
1. There is likely to be a high learning curve, particularly if the user is going to design and set up the database from scratch.
2. The initial labor for setup is intense and may require maintenance and some development work.
3. It is difficult to change the structure once the database is already in use, so the design should be airtight.

Conclusions

Whether or not a library or department officially supports a research data management service, these are actual and immediate needs of humanities scholars that can likely be addressed through existing services and expertise. Whether it is with public service outreach to graduates and undergraduates about digital information literacy, or one-on-one consultations with faculty trying to get a grasp on an image archive on their hard drive, information professionals are ideally situated to offer support and instruction on managing and organizing research data for scholars in the humanities. While there is much that we as humanities service providers could learn from the STEM services around research data management, there is just as likely to be much that STEM providers could learn from us about managing image collections.