Image Management as a Data Service

by Berenica Vejvoda[1] K. Jane Burpee[2] Paula Lackie[3]

Abstract

Across all disciplines, researchers are creating or gaining access to an ever-growing body of digitized images. Since research data management includes the organization of 'all materials' intrinsic to a research project, a robust data management plan will include a path for images as well as data in the more traditional sense. While researchers across disciplines have a long history with the organization of numeric data, the inclusion of images as a resource set in research is only starting to take shape across the disciplines. This paper is intended for data librarians or academic support staff without expertise in image data management. The primary focus is to apply traditional data management practices to images and to discuss the challenges associated with managing image collections through the research data lifecycle.

Keywords

data management, images, research data lifecycle, preservation, sharing, mixed-methods research

Introduction

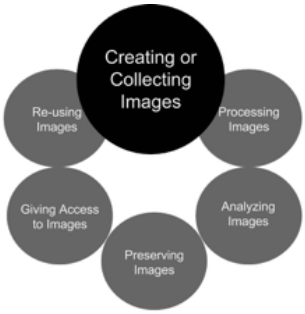
To move in small measure towards a greater understanding of image collections as a data management challenge, this article compares traditional numeric data management with organized image collections. By conceptualizing image collection management as a component of data service across the 'research data lifecycle' we hope to foster a better understanding of how data professionals can effectively transition their skills to include the management of images. This paper addresses images as data rather than as an object. The unique challenges associated with images (versus numeric data) will be highlighted through the points of the research data lifecycle which are most impactful for image management. Although the principles outlined here may apply to any digital object identified as an "image", this article will assume a format-based approach that includes any two-dimensional digital image format.

A Research Data Lifecycle Approach for Research-Related Image Collections

The following stages images in the data lifecycle (see diagram at right) will be discussed in the comparison with numeric data and research-related image collections: creating or collecting images; processing images; analyzing images; preserving images; giving access to and re-using images (adapted to images based on DCC, 2012 and MANTRA, 2014).

Adapted from, Create and Manage Data - Research Data Lifecycle. UK Data Archive, 2016. Retrieved from http://www.data-archive.ac.uk/create-manage/life-cycle. Copyright 2002 -2016: University of Essex.

Creating and Collecting Images as Data



Managing data during the creation stage of the research data lifecycle can be challenging, most notably when the 'data' are in the form of images. Images can be collected in a number of different ways, e.g.: in-house or outsourced scanning or photography; digital creation from the outset; or purchased from vendors (Primary Research Group, 2013). Just like any data gathering process, for collection methods to be successful researchers will benefit if they make decisions before they begin the process of collecting and capturing images. Therefore, questions commonly asked when

collecting numeric data offer a useful framework for effectively collecting and organizing image as a data-style resource (DCC, 2013).

- What type of data will be gathered?
- In what 'format' will the data appear? (will it change through the life cycle of the project?)
- What will be the expected 'volume' of data collected?

Types of Data ~ Types of Images

At the most fundamental level, digital images are numeric data. They are all ones and zeros stored on computer media. In practice however, images are often more like social science microdata, they can be both data as well as a datum at the same time. For example, a collection of images organized around a particular theme is comparable to a dataset and the individual image, a specific response source. When considering an image management strategy, simultaneously managing images with their metadata is like managing survey metadata, paradata, and data all at once. For example, both surveys and image collections may have metadata, paradata, and data important to the analysis or for reuse purposes.

For the purposes of data management planning, it might be helpful to think of these example parallels:

	Survey Respondent	Digital Image
Data	survey answers	Often simply the viewable image but it may also be the direct analytic content derived from that image
Metadata	e.g. age, education level, home address	e.g. camera make & model, image timestamp as set by the camera, aperture, shutter speed (EXIF data)
Paradata	e.g. respondents click-rate through a survey,	e.g. average image color, facial recognition material, image sequence in a set

working through choices in an image data management plan, comparing the process with data associated with a survey respondent can be a good place to start; both yield information elicited through data collection instruments, are inherently complex, and it is necessary to make choices regarding which aspects to focus on. Of course, the analogy is limited since unlike survey respondents, an image database may also contain the complete image - which may be flawlessly duplicated - unlike humans.

In addition to the need for understanding the complexity and structure of images as data, for images to be useful, to those who create them as well as for subsequent re-use, it is equally necessary to consider the format of image files as early as at the point of creation. Attention to format will ensure long term access and functionality.

Data Formats ~ Image Formats

Deciding on the best file format (i.e. the way information is encoded in a computer file) is understandably a question that applies to both numeric data and images. Both rely on applications or programs that will recognize the file format in order to access information within the file. According to a report published by a Digital Preservation Policy Working Group at Cornell (2001) file format for images consists of the bits that comprise the image and the header information on how to read and interpret the file. Similar to numeric data files, images can be stored in a wide variety of formats, including: bitmap (BMP), Joint Photographic Experts Group (JPEG), JPEG 2000, and Tagged Image File Format (TIFF). These standard image formats vary in relative file size, image quality and flexibility, and compatibility with software programs. Distinguishing between minimal requirements and recommended imaging requirements, the Cornell report gives preference to TIFF formats as a master image format since they do not compromise data, while JPEGs, a "lossy " format which compresses data, are included in the minimal criteria.

One research domain in particular that has championed the adoption of open standardized image data formats is the imaging community in the biological sciences, namely the JCB DataViewer initiative. Initiatives like the JCB Dataviewer align with the conventions of numeric data that recommend the adoption of open source formats in order to retain the best chance for future readability. If open source isn't an option then choosing formats that are in widespread use or agreed-on international standards will help achieve the objective of longevity and/or replicable research.

Volume: Counts and File Sizes

Related to formatting is the notion of volume. When data were first digitizable, disk memory was extremely limited. Researchers were resourceful in how they encoded and managed their data. As expectations for robust data analysis were fed by Moore's Law and a parallel rise in disk storage capacity enabled the rise in big data (e.g. moment-to-moment stock trade data or global social network data). Likewise, expectations for big data - in the form of images - has also risen. Like numeric data, a big concern for image data formats is related to a combination of storage space and computational power. Just like with numeric data; the numbers of image collections, the numbers of images in collections, and the size of individual images are all growing. Researchers should be aware of the trade-offs they are making when choosing either fewer images or images of lower quality than the original images as they were created.

Dealing with large numbers of lossless image files can quickly become unwieldy in terms of available storage space. Compressing large image files to smaller files is most easily produced using lossy compression and while the images may still provide adequate information for the immediate intended purpose, their longevity may suffer.

An advantage of the usual numeric data compression over image file compression is that they are lossless (e.g. .zip, .7z, and .gz.) On the other hand, choosing smaller image file formats (e.g. JPEGs) that are lossy over lossless image files (e.g. raw, TIFF) results in loss of image clarity; resolution, layers, and fidelity. As a result, the management of images becomes a more difficult decision when dealing with a large number of large files. In terms of image management, due to the usual lossy nature of compression, there is a clear preference for retaining uncompressed versions or for working to manage the balance of lossless compression against future format compatibility challenges. At the very least, it is recommended that researchers minimize the number of compression processes that need to be managed over the long term (Cornell, 2001).

The challenges associated with large image files are compounded by the sheer volume of images being produced across various disciplines and sectors. While the growth rate of images can be difficult to quantify and many claims appear as unsubstantiated hyperbole, the discourse surrounding the explosion of visual content agrees that it is undeniably large (Kane and Pear, 2016). Kane and Pear (2016) estimate that while 3.8 trillion photos were taken in all of human history until mid-2011, one trillion photos were taken in 2015 alone. In academia, specifically, the biological sciences, Moore, Allan, Burel, Loranger, MacDonald, Monk and Swedow (2008) noted eight years ago that 'most laboratories and imaging facilities do not have the means to store the volume of data generated by their microscopes in manageable and affordable way' (p. 557). Another testament to exponential growth in the biological sciences comes from the rapid progress in genome sequencing technology. In 2011 Gross noted that second-generation machines like Illumina's Genome Analyzer II create vast amounts of images and that the volume of these images was growing by five terabytes a day. The volume of images as data produced through medical imaging is also staggering. MarketandMarkets (2016) estimate that medical image archives are increasing by 20-40% annually, and they predicted that by 2012, there will be 1 billion medical images stored.

Processing Images as Data

The growth of digital images in size and number, the advent of powerful digital cameras and the willingness of libraries and archives to use them, has produced an overwhelming need for comprehensive image management software (Roy Rosenzweig Center for History and New Media, 2016). This need is documented across disciplines by researchers who struggle to manage collections of digital images. In response to this need, in 2015, the Andrew W. Mellon Foundation announced funding for a new project to develop Tropy an open-source software application that will help researchers collect and organize digital photographs, create metadata, and export photographs and associated metadata to other platforms (Centre for History and New Media, 2016).

Researchers in the sciences – particularly, in the medical and life sciences, are also expressing a need for image management systems. The Open Microscopy Environment (OME) Consortium has, for example, built a series of open source tools that assist researchers in managing large sets of complex images to support research in cell and developmental biology (Moore et. al., 2008). Researchers relying on medical imaging (e.g. CT, MRI, X-ray, NM, mammography, ultrasound, radiology) are also in need of image management systems to keep up with unprecedented growth. A case in point comes from the critical role of medical imaging, specifically image biomarkers, in clinical trials for Alzheimer's disease. Increased reliance on these medical images for study outcomes requires image management systems for effectively capturing, processing, analyzing, disseminating and archiving images (Jimenez-Maggiora, Thomas, Brewer, Bruschi, Hong, and Aisen, 2012). Jimenez-Maggiora et. al. (2012) note that these specialized systems are complex, inflexible and resource intensive.

Recommendations for managing traditional numeric data files at this stage of the research data lifecycle can provide a useful framework whether the data are "Big" or just awkward. Key considerations for organizing numeric data include data carpentry functions, such as: versioning, naming, and renaming (MANTRA, 2014). As with a numeric data file, an image file name needs to be carefully considered for consistency, logic and predictability so that users can effectively browse and retrieve image data and also to avoid confusion when multiple researchers are working and naming shared files. In other words, 'ThisImage.tiff' will be as problematic as 'ThisSASFile.sav'. And similarly, images will present with multiple files in various formats, multiple versions, across differing methodologies, etc.

A growing number of software tools exist to help organize images in a consistent and automated way through functions such as batch renaming. Renaming files may be especially useful for image metadata in instances where digital cameras automatically assign base filenames of sequential numbers. In the realm of numeric data file management, tools include: using the GREP command in UNIX or applications such as RenameIT . In addition, ImageMagick can perform various batch processing functions on image metadata. In addition to batching renaming, image management software can support image workflows by assisting with recording location, generating thumbnails and storing basic associated metadata that are embedded in image files (Jisc, 2016).

Analyzing Images as Data

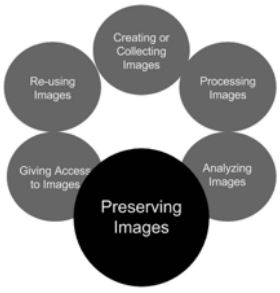
One of the defining features of data is that they are the raw material produced by primary research that is intended for analysis (Geraci, Humphrey, & Jacobs, 2012). Arguably, numeric research data are most often created for the intended purpose of analysis. In contrast, images may not always be intended for analysis at the point of creation. For example, many early digital library projects supported the creation of images to add to library collections, but in a large majority of cases such images were and continue to be used by researchers for making examples and illustrations versus serving as raw material for research. It is important to note that an image or collection of images may serve a variety of users and as such, they may also become data for analysis.

Recommended best practices for managing numeric data at the analysis stage of the research data lifecycle (MANTRA, 2014; DCC, 2012) include documenting analyses and file manipulations, managing versions of data files and deciding if analyzed data will be shared. With numeric datasets, it is fairly routine to document analysis and manipulation functions, usually in the form of programming code files. Similarly, documenting the manipulation and analysis of images helps researchers with their own image processing and analysis workflows (e.g., logging the numerous steps taken to geo-reference an image) and will also produce greater transparency of techniques, critical to successful replicability, which have in recent years emerged as potential sources of controversy across many disciplines. According to McCook (2016), companies such as Image Data Integrity (IDI) exist to help journal editors, publishers, funding agencies and institutions screen and verify whether image manipulation (e.g., blots and micrographs in biomedical materials) compromises the interpretation of the images for scientific purposes. In addition to documenting the process of analysis and manipulation,

researchers working with numeric data also decide what form of data to ultimately share: i.e. raw, processed, analyzed, final.

The analysis stage of the research data life cycle presents unique challenges for those managing image data when using image analysis software tools. By way of example, ImageJ , which has existed for over 30 years, is a general-purpose, extensible scientific image-analysis program that is used to capture, display and enhance images in the biological sciences (Schneider, Rasband & Eliceiri, 2012). One key challenge noted by Schneider et. al. (2012) occurs when one uses the software to open and parse the countless variety of image file formats. In the case of proprietary file formats tied to specific software (e.g. with some microscopes), using such software can be especially problematic for reproducibility or any further sharing. It is preferable to have images from research processes be available independently of specialty equipment. ImageJ, for example, is able to connect directly to MatLab so that researchers are able to run statistical analyses as well as other tools such as Imaris which supports 3D and 4D image analysis (Schneider, et. al., 2012). In order to encapsulate diverse needs even within a particular domain w biological imaging, image analysis tools need to remain flexible and extensible.

Preserving Images as Data



One of the most critical aspects of data management, regardless of data type is the pairing of metadata to accurately and sufficiently describe datasets. It is important to note that like most other data management functions, metadata hold a central role across the entire span of the research data lifecycle. So while it is discussed in reference to preservation for the sake of structuring this discussion, it applies to other stages as well. Metadata for images is especially critical as a way to organize and search through growing libraries and repositories of images that are being produced by researchers and consumers alike.

As is the case with numeric data files, decisions need to be made about how much detail to record in image metadata records. The internationally accepted metadata standard for describing social science numeric data, DDI (Data Documentation Initiative) can be crudely parsed between study-level descriptions and variable-level metadata. This distinction can also apply to images. For example, low level descriptors such as 'title', 'creator', and 'size' are similar to DDI fields such as 'title', 'abstract', 'producer', 'distributor', and 'time period'.

For more detailed descriptions, DDI offers metadata fields at the variable-level – e.g. exact meaning of the datum (ICPSR, 2016). This level of description is created directly from formatted datasets. While the need for fuller descriptions of image content are also equally necessary, a key difference is the degree of subjectivity involved in producing more abstract, higher level meaning, such as feelings portrayed by a particular image, 'happy', 'sad' (Jisc, 2016). This challenge for describing images is discussed extensively by Eadie (2008) in his description of the development of the Jisc-funded Dublin Core Images Application Profile. Eadie (2008) aptly notes that unlike text-based materials that can be machine processed, images are not easily self-describing. This reasoning can also contrast with numeric data where machine-processing can quite easily produce meaningful variable and file-level metadata based on objective information embedded in formatted statistical data files. Digital images, on the other hand, have a more complex relationship with machine-assisted metadata-extraction; while pixel data and bit depth are objective data points, they provide little by way of meaningful contextualization of images as data (Eadie, 2008). So despite increased quality and quantity of camera sensor elements, it is neither practical nor meaningful to describe images to aid organization or querying based on millions of image pixels (Metadata Group, 2008).

In addition to subjectivity, images can be complex to describe because they often have relationships with other objects - which may even be embedded within them. So before you can even begin to say anything about an image you need to be very clear about which aspects of it, or its relationship with other objects, on which you are actually focusing (Jisc, 2016). For example, images can be found in slides, photographs, books, manuscripts, lectures, and presentations. There may also be interdependencies between images. For example, in the area of Geographic Information Systems (GIS), different layers of GIS data are superimposed to create a richer representation for spatial analysis.

Adding to the complexity of image description is that description consists of at least three different types (Eadie, 2008). First, there is technical information relating to the image. This is usually pretty straightforward as capturing technical metadata is usually automated (e.g., captured by digital cameras) and resides in the image itself. The main metadata container formats for images are: Exif (Exchangeable Image File Format) (for device properties), IPTC (International Press Telecommunication Council), IIM (Information Interchange Model) (workflow properties), and Adobe XMP (Extensible Metadata Platform). Each metadata container format has unique rules regarding how metadata properties are stored, ordered and encoded (Metadata Working Group, 2010). While technical metadata are fairly easily captured, how they are structured is considerably more complex. Even within the container format, metadata are stored, for example, according to various semantic groupings; within these groupings there can be numerous individual metadata properties. Perhaps the biggest issue concerning the structural complexity of technical metadata is that different applications and devices handle these technical specifications in different ways; hence, creating challenges for interoperability. Technical metadata also becomes more complicated for long term preservation as more metadata fields need to be added that are not normally captured by devices – for example, image format migration and versioning.

The second and third types of description for images as data relate to the content in an image; the application of abstract principles to the description of the image. Not only are content and abstraction difficult to describe in a standard way, but text-based descriptions will vary depending on the knowledge, culture, experience and point of view of the cataloguer (Jisc, 2016). Even more difficult is anticipating the needs of users in terms of what to describe in images. For example, in Figure 1, one researcher may be drawn to the couple while another, looking for depictions of leisure activities would find the hoop-rolling relevant. One way to deal with this challenge is to balance the time and resources available for describing images with the anticipation of what level of detail users of the image will require for effective discovery (Jisc, 2016). Given the inherent subjectivity and richness of images, however,

rarely in large image libraries and archives are there sufficient resources for in-depth description.

In addition to the critical role of associated metadata, image preservation involves decisions about where and how to store them. As is the case for numeric data, depositing images in an archive or repository should facilitate discovery and preservation for the long-term. As mentioned earlier, some domains such as genomics produces such vast quantities of images (more than five terabytes a day; see Gross, 2011) that the practicality and cost of archiving these images for preservation purposes is often not feasible. Gross (2011) instead points to a motivation to alternatively invest in the development of real-time processing of the images 'to output only the base calls and the quality values' (p. R204). The distilled nature of the images for their originally intended purpose will limit future re-usability of these images but current technical and financial limitations mandate that some hard decisions are being made regarding precisely what to preserve.

With the currently overwhelming volume of images as data we see additional kinds of re-use obstacles; image collections constitute "big data" in that they are larger than can currently be managed, adequate storage space is another issue, and even if storage were available, the transfer of such large files or sets of files is currently impractical. This should however be prefaced with a note that new infrastructure initiatives such as the pan-European project called ELIXIR (European Life Science Infrastructure for Biological Information) are looking for solutions that balance software compression with judicious data reduction. With ELIXIR, the ultimate aim is to bring compression down to 0.1 bits (0.01 bytes) for every base stored - which translates to a human genome taking up just 30MB of storage (Gross, 2011) (as opposed to the 1.5GB without the compression).

Not all image collections are so unwieldy. For the more common, manageable collections of images, it is possible to decide where to best archive image sets and their metadata. Because images as data are produced across a vast number of domains, repositories can range from individual solutions for photographers and other artists, to institutional photographic and slide collections, to archives and museum image repositories, and as institutional teaching and research archives.

As previously mentioned in reference to open/standardized data formats, JCB DataViewer was the first open repository in the life sciences that allowed for archiving and sharing of original image datasets to support published scientific articles (Linkert, Rueden, Allen, Burel, Moore, Patterson, Loranger, Moore, Neves, MacDonald, Tarkowska, Sticco, Hill, Rossner, Eliceiri and Swedlow, 2010). In addition to archiving the original binary image and associated metadata, additional information captured by acquisition software includes: acquisition settings, image size, and resolution.

While the imaging community in the life sciences already treats images as data and wherever possible has robust archiving solutions, this is not the case in all areas of research where digital images are produced. For these areas, the consideration for archiving numeric data can provide guidance for treating images as data in need of long-term preservation. Arguably, university repositories that have traditionally focused on archiving text-based resources may benefit from examining numeric data archiving practices as they can assist in archiving images as data.

As we know, preserving numeric data can be problematic due to the amount of data being generated (MANTRA, 2014). Given that image files are substantially larger, on average, this becomes a key consideration. Additionally, reliance on specific technologies for accessing anything digitized - becomes problematic since the technologies change quickly. Therefore, as with numeric data, it is essential to archive digital image data in a systematic way in order to minimize the chance of obsolescence or making images inaccessible over the long term.

Accessing and Sharing Images

The sharing and accessing phase of the data lifecycle is perhaps where those working with images become easily overwhelmed and/or frustrated when a discrepancy emerges between needs and search results (Chung & Yoon, 2011). As noted in other phases of the lifecycle this challenge is exacerbated by the explosive growth and availability of images. In order to allow for effective access to images, the growing image collections must be organized in ways that allow for efficient discovery, browsing, searching and retrieval (Rui, Huang & Chang, 1999).

According to Wang, Mohamad & Ismail (2010), an effective image retrieval system needs to be able to retrieve relevant images based on queries that conform as closely as possible to human perception. So unlike quantitative numeric data files, which are relatively straightforward to describe based on keywords that map to the represented measures, visual information is far more ambiguous and semantically rich (Wang & Ismail, 2010). Relying on traditional keyword querying systems of access will not be sufficient. Wang et. al. (2010) Note that image retrieval based on keyword querying, popular in the 1970s, relies on keywords used as descriptors to index an image. While the Jisc Digital Media Guide (2016) discusses the need for advanced search features (e.g. Boolean logic) to support relevant keyword image retrieval, Wang et. al. (2010) would argue that assigning keywords manually to images is not only time consuming but that keywords alone are inadequate and grossly inefficient to describe the rich content of images.

Another trend in image retrieval that supersedes text-based image retrieval (popular in the 1980s), is content-based image retrieval (CBIR). First used by IBM, this method of retrieval is based on extracting visual features from the image itself. While in theory CBIR systems can include the extraction of low or high level features, even with sophisticated algorithms that can combine multiple visual features, elements such as colour, texture, shape and spatial relationships do not come close to mirroring the richness of image content that users have in mind when searching for relevant images.

As a proposed solution, Wang et. al. (2010) discuss at length the most recent trend of developing semantic-based image retrieval systems that allow users to query image data using high-level concepts. In short, this retrieval method maps the automated process of extracting low level visual features from images with semantic descriptions also stored in an image database. It is hoped

that this example of intelligent image retrieval, that can better represent the abstract concepts inherent to images, will help users discover and access relevant images.

While semantic-based image retrieval, in theory, allows users to better discover relevant images based on higher-level meaning, creating semantic descriptions still requires human intelligence which is time consuming and expensive. One innovative way to tackle this dilemma is to consider the need for metadata creation by users during the access and reuse phases of the research data lifecycle. Numerous web-based initiatives are testament to this type of metadata crowdsourcing (or 'social tagging'). ArtUk, an online site for art from every public collection in the UK, recently launched, ArtUK Tagger , which allows the public to add multiple tags to paintings. An algorithm then calculates which tags are likely most accurate and feeds these tags through to the Art UK website. Similarly, the Philadelphia Museum of Art encourages online visitors to tag objects in the online collection in order to improve access to works of art. Social tagging initiatives not only exist to provide better subject access to images but also to assist with quality control and processing functions. MicroPasts - for example, encourages users to help with location accuracy of artifact findspots and photographed scenes as well as the masking of photos intended for 3D modelling. Zooniverse is yet another platform that is designed to use volunteers to sort through and help classify excessive numbers of research images. "Our goal is to enable research that would not be possible, or practical, otherwise."

While crowdsourcing initiatives and semantic-based search functionality can improve access to relevant images, building an image database based on shared standards remains a challenge given the diversity of image collections, widely varying budgets and differing individual requirements of user groups (Bourne, 2005). Regarding the varying requirements of users, Chung and Yoon (2011) found that users were more likely to search based on abstract meaning when images were intended as objects but not when used as data. Another constraint that has similarly plagued the management of numeric data collections in university settings is that images (particularly slides) generate data that are very different from those handled by cataloguing systems created for books (Bourne, 2005).

Re-using Images

In addition to browsing, search and retrieval challenges associated with managing access to images, a discussion of the closely associated re-using phase of the data lifecycle is not complete without attention to image copyright as well as issues regarding confidentiality or data sensitivity. Generally speaking, factual numeric data in and of itself, represented in an obvious file structure, is not copyrightable in Canada as a work needs to be original for copyright to exist. This holds no matter how much work goes into collecting the data (Potvin, 2008). However, most numeric data need to be analyzed and processed and so the program code developed for these purposes is considered a 'literary work' (Potvin, 2008). Also, once data are formatted into, for example, a relational database, graph or dataset, it can be subject to copyright.

As previously mentioned, images, unlike traditional numeric data files, are not always created for the purpose of data analysis. As such, most literature discussing images and copyright reference images as artistic works that are copyrightable. Images considered to be artistic works in the UK, for example, include: blueprints, building plans, cartoons, charts, decorative graphics, diagrams, drawings, engravings, graphs, illustrations, logos, maps, moving images, paintings, photographs, sculptures and sketches. According to the Government of Canada images as artistic works include: patterns, art slides, maps, paintings, architectural drawings, plans, digital images, drawings, photographs, charts, and art prints.

Just as with copyrightable numeric datasets, it is necessary to clarify who has primary ownership of the datasets since copyright of a work comes into existence at the point of creation (i.e., authors/ creators). If images are treated as data then similar ownership and rights issues apply when figuring out how the images will be managed and disseminated. This will mean assessing whose rights need to be considered, for example: funders, institutions, research participants, collaborators, publishers and the public (MANTRA, 2014).

Copyright is undeniably complicated and there are obvious, notable exceptions to the principle of creator as copyright holder. For example, U.S. federal copyright law denies copyright protection for works produced by the US federal government so NASA's images, for instance, are in the public domain, but individuals who create images based on data released by NASA can assert limited copyright because they have created derivative works or compilations.

Additionally, copyright surrounding images is context-specific. Images may simply exist as facts (equivalent to a numeric data file), in which case, they are not subject to copyright. Copyright may also not be an issue if copyright has expired or images are considered to reside in the public domain. Some image owners may also allow reuse for non-commercial purposes (i.e., education) but require attribution (e.g., Creative Commons Attribution). The educational sector may also find that images fall under fair use or fair dealing. In support of this, the Visual Resources Association (VRA) in the U.S. published a statement on the fair use of images for teaching, research and study (Wagner & Kohl, 2012). For teaching and study purposes, this statement covers preservation, use (both high-resolution and thumbnails), adaptations, sharing and reproduction. Also, if images are photographs, they are likely to be treated as original works and subject to normal copyright restrictions. Collections of images may be copyrightable if they exist as a database or as a result of researchers creating added value to images.

Ethical considerations, specifically privacy and confidentiality, also merit careful consideration when managing and sharing images. As with numeric datasets, researchers managing and disseminating images need to minimize the risk of disclosing confidential information and re-identifying study participants. One broad technique for safeguarding confidentiality of numeric data includes either collecting data without identifiable information or anonymizing data post-collection through de-identification processes. Best practices for handling sensitive image data may include the anonymization of facial and location identifiers in digital photographs. The actual techniques for doing so, however, may pose unique challenges for images. For example, in a 2016 email thread on the Jisc Research Data Management Listserv, the

issue of anonymizing image data proved labour intensive and expensive for a use case where anonymization was required for a large collection of images. They couldn't find a freely available tool that could effectively bulk-blur identifying characteristics in the images. In another comment from the same thread, a researcher noted that obscuring faces by pixelating sections of a video image could greatly compromise the usefulness of data. Alternative strategies to anonymization noted by many researchers are to either gain consent to share, or to consider controlled access so that the usability of the images can remain unaltered. To maximize the effectiveness of these alternative strategies to anonymization, strong recommendations are made to consider and judge at an early stage the implications of depositing images with confidential information.

Figure 1

Source: LeBlond & Co. Her Majesty at Osborne. Regal series ca 1850. Print collection, Rare Books and Special Collections, McGill University.

Conclusion

The unprecedented growth of research image collections across disciplines, coupled with increasingly powerful instruments and devices for image capture, have created challenges and new opportunities for managing images across the research data lifecycle. This paper offers some preliminary recommendations for managing images as data by looking to established research data management practices for traditional numeric datasets.

Analogously, images, like survey respondents will provide information in the form of data. But like people, images are inherently richer than the discrete slices of data that are extracted by research instruments. Uniquely, images pose challenges in terms of size and volume, especially for storage and preservation. The creation of robust metadata is also complicated due to the subjectivity of image meaning and the difficulty in anticipating the search needs of users. Also, automating the process of metadata creation is difficult and manual description remains necessary. Some emerging solutions to these challenges are noted, including crowdsourcing initiatives for data processing and description as well as semantic-based retrieval systems.

References

Avondo, J. (2010) BioformatsConverter. (Available at cmpdartsvr1.cmp.uea.ac.uk/wiki/BanghamLab/index.php/BioformatsConverter)

Bourne, M. Image data. VRA Bulletin, 31(3), 26-29. (Available at http://web.b.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=7&sid=106aa9eb-e435-4f6f-a71e-01e5c781e0f3%40sessionmgr106&hid=107)

Center for History and New Media (2016). RRCHNM to build software to help researchers organize digital photographs. (Available at https://chnm.gmu.edu/news/rrchnm-to-build-software-to-help-researchers-organize-digital-photographs)

Chung, E. and Yoon, J. (2011). Image needs in the context of image use: An exploratory study. Journal of Information Science, 37(2), 163-177 (Available at http://jis.sagepub.com/content/37/2/163.short)

Corti, L., Van den Eynden, V., Bishop, L., and Woollard, M. (2014). Managing and sharing research data: A guide to good practice. Los Angeles: Sage Publishing,

Humphrey, C. (2006) e-Science and the Life Cycle of Research. (Available at http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc)

DDI Alliance (2013) Data Documentation Initiative. (Available at http://www.Ddialliance.org)

Data Curation Centre (DCC) (2012). (Available at http://www.dcc.ac.uk/resources/curation-lifecycle-model)

EDINA. (2014). MANTRA (Available at http://datalib.edina.ac.uk/mantra)

Eadie, M. (2008). Towards an application profile for images. ARIADNE: Web Magazine for Information Professionals. http://www.ariadne.ac.uk/issue55/eadie

Geraci, Humphrey, & Jacobs (2012). Data Basics: an introductory text. Unpublished. Local PDF file.

Gross, M. (2011). Riding the wave of biological data. Current Biology, 21(6), R204-R206. (Available at http://ac.els-cdn.com/S0960982211002818/1-s2.0-S0960982211002818-main.pdf?_tid=e46374de-026b-11e6-b386-00000aab0f26&acdnat=1460657489_27698d6d824b844bbc3b43f55c85558e)

ICPSR (2016) Data Management & Curation: Metadata (Available at https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/lifecycle/metadata.html)

Jimenez-Maggiora, G. A., Thomas, R. G., Brewer, J., Bruschi, S., Hong, P., and Aisen, P. S. ADCS Electronic Data Capture (EDC) - Integrated Multi-Modal Image Management for Clinical Trials in Alzheimer's disease. Neurosciences Department, University of California at San Diego: La Jolla, CA,(Available at https://www.researchgate.net/profile/Gustavo_Jimenez-Maggiora/publication/269038101_ADCS_Electronic_Data_Capture_(EDC)_-_Integrated_Multi-Modal_Image_Management_for_Clinical_Trials_in_Alzheimer's_Disease/links/548734e30cf268d28f071e7c.pdf)

Jisc Digital Media (2016) (Available at http://www.Jiscdigitalmedia.ac.uk)

Kane and Pear (2016) (Available at http://sloanreview.mit.edu/article/the-rise-of-visual-content-online)

Linkert, M., Rueden, C.T., Allan, C., Burel, J.M., Moore, W., Patterson, A., Loranger, B., Moore, J., Neves, C., MacDonald, D. and Tarkowska, A., (2010). Metadata matters: access to image data in the real world. The Journal of cell biology, 189(5), 777-782.

MarketsandMarkets. (2016). Rising volume of medical imaging data to increase the adoption of cloud computing in the healthcare sector. (Available at http://www.marketsandmarkets.com/ResearchInsight/north-america-healthcare-cloud-computing.asp)

McCook (2016) Retraction Watch. Don't trust an image in a scientific paper? Manipulation detective's company wants to help. Retraction Watch. (Available at http://retractionwatch.com/2016/02/24/dont-trust-an-image-a-new-company-can-help)

Metadata Working Group. Guidelines for handling image metadata. (2010). (Available at http://metadataworkinggroup.com/pdf/mwg_guidance.pdf)

Moore, Allan, Burel, Loranger, MacDonald, Monk and Swedow (2008). Open Tools for Storage and Management of Quantitative Image Data. Chapter 24 (Available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.461.9978&rep=rep1&type=pdf)

Potvin, J. (2008). How is copyright relevant to source data and source code? Technology Innovation Management Review. (Available at http://timreview.ca/article/121)

Primary Research Group. (2013). Survey of Best Practices in Digital Image Management. New York: Primary Research Group.

Keeney, A. R. and Rieger, O. Y. (2001). Report of the Digital Preservation Policy Working Group on Establishing a Central Depository for Preserving Digital Image Collections. (Available at https://www.library.cornell.edu/preservation/IMLS/image_deposit_guidelines.pdf)

Research Data Canada, (2014) (Available at http://www.rdc-drc.ca)

Schneider, C. A., Rasband, W. S., and Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nature Methods, 9(7), 671-675. (Available at https://www.researchgate.net/profile/Kevin_Eliceiri/publication/228085958_NIH_Image_to_ImageJ_25_years_of_image_analysis/links/0fcfd4fee589e852eb000000.pdf)

Statistics Canada. (2015). Section 4: Data. (Available at http://www.statcan.gc.ca/eng/dli/guide/toc/3000276)

UKDA (2011) (Available at http://www.data-archive.ac.uk/media/2894/managingsharing.pdf)

Wagner, G. and Kohl, A. (2011). Visual Resources Association: Statement on the fair use of images for teaching, research and study. VRA Bulletin, 38(1), 1-18.

Wang, H. H., Mohamad, D., and Ismail, N. A. (2010). Toward semantic based image retrieval: review. Proc. SPIE 7546, Second International Conference on Digital Image Processing, 754626 (Available at doi:10.1117/12.853332).

Notes

1. Berenica Vejvoda (MISt, University of Toronto) is a data librarian at McGill University. Prior to McGill, Berenica worked as a data librarian at the University of California at San Diego and at the University of Toronto. berenica.vejvoda@mcgill.ca
2. K. Jane Burpee (MLIS, McGill University) is the Coordinator, Data Curation and Scholarly Communications at McGIll University. She has been active in the area of scholarly communication since 2000 when she became a scholarly communication librarian at the University of Guelph. She is a leading voice for open access and champions the transformation of scholarship. jane.burpee@mcgill.ca
3. Paula Lackie (MA, A.B.D., University of Southern California) is the Academic Technologist for Data at Carleton College. She is long-time research data advocate and social science and humanities technologist. plackie@carleton.edu
4. The Research Data Lifecycle (Humphrey, 2006; DCC, 2012; DDI Alliance, 2013, MANTRA, 2014).
5. Lossy compressions transform and simplify the media information in a way that gives much larger reductions in file size than lossless compressions. While the file becomes significantly smaller, quality of the image is compromised during the compression process (e.g. JPEG). Conversely, lossless compression results in no information loss, however, the image files are much larger (e.g. TIFFs) Jisc, 2016 (Available at http://www.Jiscdigitalmedia.ac.uk/infokit/file_formats/lossless-and-lossy-compression)
6. JCB DataViewer (https://datahub.io/dataset/jcb-dataviewer), launched in 2008, for archiving and sharing original image data in the life sciences, allows users to download original image data in an open, standardized data format and preserves the original image metadata (OME tagged image file format [TIFF]). Similarly, the Jisc-funded Data Management for Bio-Imaging project at the John Innes Centre developed BioformatsConverter software (Avondo, 2010) to batch convert bio images from a variety of proprietary microscopy image formats to the Open Microscopy Environment format, OME-TIFF. OME-TIFF, is an open file format that enables data sharing across platforms and maintains original image metadata in the file in XML format (UKDA, 2011).
7. Moore's Law refers to the long-standing pattern that computer processing power will double every two years.
8. Big data is currently an ill-defined term that at its root simply refers to extremely large data files that require greater than average computational power to manipulate and/or analyze.
9. Tropy: http://chnm.gmu.edu/news/rrchnm-to-build-software-to-help-researchers-organize-digital-photographs
10. RenameIT https://github.com/wernight/renameit
11. ImageMagick: http://www.imagemagick.org
12. A "thumbnail" is a very small version of the original image. Thumbnail versions are useful as a kind of wordless summary of the image.
13. ImageJ: https://imagej.nih.gov/ij/
14. ArtUK http://artuk.org/tagger
15. Micropasts Crowdsourcing: http://crowdsourced.micropasts.org
16. Zooniverse is a "platform for people-powered research": https://www.zooniverse.org The development of which is funded by generous support, including a Global Impact Award from Google, and by a grant from the Alfred P. Sloan Foundation.