Published in final edited form as:
Mol Reprod Dev. 2015 September ; 82(9): 630–634. doi:10.1002/mrd.22538.

Data Management and Archiving in a Large Microscopy-and-Imaging, Multi-User Facility: Problems and Solutions

CALLEN T. WALLACE, **CLAUDETTE M. ST. CROIX**, and **SIMON C. WATKINS**[*]
Center for Biologic Imaging, Department of Cell Biology, University of Pittsburgh, Pittsburgh, Pennsylvania

SUMMARY

Advancements in microscopy and imaging have pushed the boundaries of what was once thought possible in many fields of research. New techniques, coupled with the application of new technologies, allow researchers to answer increasingly complex questions by probing deeper and with greater accuracy. While, these new techniques provide far greater specificity and increased sensitivity in regards to both resolution and frequency, the amount of data generated is swelling to a point where conventional data-management systems struggle to keep pace; this is especially true for large microscopy-and-imaging shared-user facilities. Sub-optimal data management can severely hinder the ability of a researcher to determine experimental results accurately or efficiently, and will inevitably limit the functionality of the research facility itself. This review discusses the source of the problem: how data are produced by systems available today, and the information's specificity and relative importance; techniques for management of these data to maximize functionality of the facility; and practices that can be detrimental in the research core environment.

INTRODUCTION

Before the late 1980s, microscopy was primarily considered a descriptive technology, and was used purely to define cellular or histological structures using basic approaches, or antibody localization with colorimetric or simple, single-color fluorescence readouts. During that period, film was the substrate-of-choice for archiving almost all microscope-derived data, the darkroom was ubiquitous, and quantitation of any structural change was extremely laborious and inaccurate. For example, the relative size of a structure was assessed by printing the image on photographic paper, cutting out the structure, and weighing it.

The capabilities of microscope-based imaging technologies grew exponentially following the development and commercial release of the confocal microscope (in 1987), the discovery of green fluorescent protein (GFP), and advances in digital camera architecture. The confocal approach presented the first potentially quantitative application of microscopy, particularly for fluorescence readouts. Although, the first confocal microscopes preceded the use of digital-imaging techniques by more than a decade, a synergistic relationship quickly

[*]Corresponding author: Center for Biologic Imaging Department of Cell Biology University of Pittsburgh, BSTS 225 3500 Terrace St. Pittsburgh, PA 15261. ; Email: swatkins@pitt.edu

took hold between the two common digital approaches of that era: utilizing either video cameras, which allowed “on chip” integration followed by conversion to a digital signal, or early charge-coupled device (CCD) cameras, based on the Kodak KAF1400 chip. While, CCD devices were extremely expensive (over $25,000 in 1991) and slow (500 kHZ), they did offer 12- or 16-bit imaging capabilities that vastly exceeded the 8-bit format of video images. Computer memory was extremely expensive at this time and bus architecture was quite slow (ISA standard), thus, the initial development of true digital imaging was limited. In 1992, however, Sony Corporation released the first interline CCD chip with microlenses. Earlier, CCD devices had a single output point, which gave rise to significant thermal issues that demanded the development of elaborate Peltier-based cooling strategies as well as vacuum protection of the CCD chip to reduce thermal noise – both of which added significantly to the cost of a camera. In contrast, the interline chip had multiple read points, so less processing was done on-chip, and hence, less thermal noise was present in the system. This lowered camera cost significantly and increased read speeds from the chip (10–30 mHz). Meanwhile, the advent of the PCI bus sufficiently increased processor speeds to deal with higher data output. The slow and steady decline in the cost of memory chips also allowed for the integration of more processing capacity, thus, enabling real-time viewing and manipulation of images.

The status quo for data complexity remained stable for about 15 years, even though the science performed with a microscope continued to evolve with the increasingly sophisticated questions being asked. Approaches transitioned from two to three dimensions, and/or from one fluorophore with a contrast image (e.g., phase or differential-interference contrast) to three or more fluorophores with a contrast image. Implementation of integrated systems that are capable of imaging live cells or animals has become increasingly common over the last 10 years. Indeed, there has been a complete revolution in the capabilities of light microscopes—which was accompanied by a rapid rise in data output and storage needs that are potentially 10 times what they were 5 years ago.

This revolution in microscopy was made possible by the development of low-cost, low-noise, complementary metal oxide sensor (CMOS) cameras, and truly fast-switching diode-based light sources. Unlike the best CCD cameras, current third-generation scientific CMOS (sCMOS) cameras are capable of collecting large (2,500 × 2,500 pixel) images with very low noise (1.03 electrons/pixel/second) at 100 frames/second with 16-bit bit depths. By comparison, point-scanning confocal microscopes still generate, at most, a single multicolor megapixel image per whereas, live-cell, multi-pinhole confocals generate fifty 16-bit 512×512 pixel images per second—both of which are quite slow. The speed of multicolor image acquisitions was accelerated closer to real-time rates by the rise-time of diodes. Traditionally, mechanical filter wheels were used to change the wavelength of the illuminating light for multicolor applications—a process that limits the speed of data acquisition, given that the average time for a filter wheel to change positions is about 50 msec; in contrast, new diode illuminators require only 0.25 msec to accomplish the same task. Thus, mechanical lag times do not limit temporal data production in multicolor applications.

Incorporation of the latest imaging systems, cameras, and techniques, however, has had a substantial impact on data production–a sCMOS camera can easily generate 1.6 gigabytes of data per second—and ultimately on the functionality of a laboratory that relies on microscopy. Consequently, data management is crucial to the successful operation of an imaging facility, and inevitably demands careful evaluation and discussion.

PARTS OF THE WHOLE: SOURCES OF BIG DATA

Bit Depth

Bit depth reflects the number of gray scales in an image. A bit has two states (on and off), hence two grayscales; therefore, a single-bit image can be considered a binary representation. Most "documentation" images have 256 grayscales, or an 8-bit depth (2^8) in each color (a standard RGB image is a combined red, green, and blue image of 8-bits per color). The 8-bit depth historically derives from the standard computer "word", or byte, which is 8 bits long. Any image with a bit depth greater than eight requires two bytes per pixel, thus, every incremental increase in bit depth above eight immediately doubles (or more) data storage or throughput needs.

When choosing which bit depth to use while imaging, it is best to consider the ultimate function of the collected image. The full capacity of a standard sensor (CCD or sCMOS) is about 20,000 photoelectrons, or about 4 photoelectrons per grayscale; few, if any, biomedical research questions demand degrees of subtlety in measurement that are greater than this. For documentation and localization information, 8 bits per channel is absolutely sufficient as the human eye can rarely distinguish more than 64 grayscales. For quantitation, a 12-bit, or 4,096-grayscale, image is the best compromise between the sensitivity of the required measurement and the file size: a single, 12-bit, three-color quantitative fluorescent image captured using a sCMOS camera is about 37.5 megabytes in size. These general guidelines apply to both wide-field and point-scanning microscopy.

Sampling and Quantitation

Most researchers have an inexorable tendency to oversample, or fail to consider the requirements for quantification when collecting data, which can result in substandard quantitation. Image collection and quantitation is readily defined by the Nyquist criterion, which states that sampling frequency should be 2.3 times the frequency of the measured event. This rule applies to both temporal and spatial images. Temporal oversampling is extremely common, and may be of little consequence, depending on the experiment, and experience with the technique. Spatial oversampling, on the other hand, is dependent on the size of the objects of interest. For example, it is not necessary to have sub-micron resolution to measure the number of nuclei (at roughly 10 microns each) in a sample; a wide-field fluorescence system equipped with a standard 10× objective is perfectly sufficient for this purpose. On the other hand, sub-micron resolution is essential when quantifying mitochondria size or endosome number, and usually requires an objective with a higher numeric aperture (NA) and magnification in conjunction with confocal methods for clear resolution of the objects. Yet, there is no point in sampling the image at a frequency that exceeds 2.3 times the resolvable limits (defined as 0.61 λ/NA in xy and 1.77 λ/NA^2 in z, where λ is the wavelength used for imaging) of the objective used.

Sample size and oversampling has become an increasing problem as multi-field, stitched images are obtained using microscopes with automated stages that can scroll through samples in the x-, y-, and z- planes. These combined images can be composed of hundreds of individual fields, and result in massive data files—for example a simple 10 × 10 field, three-color, 12-bit image will be 3.75 gigabytes. Such a data scale is reaching the point of becoming unwieldy, resulting in an increase in image load times, processing times, and sub-sampling times from many second to minutes.

Image Compression and Quantitation

Cameras today, are exponentially more sensitive to grayscales than the human eye. This increased mechanical sensitivity has both positive and negative consequences: Greater sensitivity allows you to observe more subtle differences, but also results in larger and more cumbersome datasets. In an attempt to make the data more manageable, however, individuals may inappropriately utilize methods of image compression without being fully aware of the degree to which they are altering their data. This perception is a consequence of the limitations of the human eye, which “tell” a user that the image has not changed after applying the compression algorithm.

Bit-depth compression is the conversion of a larger bit depth to a smaller one, such as 12- to 8-bit. The resulting image may seem to have little to no visual differences, depending on the display capabilities of the monitor as well as that of the graphics card used to view it, which leads to the misconception that quantification of these images will result in the same measured accuracy and specificity. The difference in grayscale range between a 12- and 8-bit image (4,096 versus 256 grayscales, respectively) is not simply a matter of more values, but lies in the ability of the camera to detect more subtle differences in the number of photons that contact the chip. Bit-depth compression from 12- to 8-bit permanently rescales the 4,096 unique values of the 12-bit image into 256 values, resulting in a 16-fold decrease in measurement specificity.

Most image-compression algorithms flatten similar regions in an image, thereby, generating large megapixels by decreasing local bit depth. The large megapixel is created by regularly sampling a number of pixels, and then reassigning the average value of the sampled pixels to the single larger pixel—for example binning. Such compression algorithms decrease the number of pixels available for measurement in the image, and thereby decrease the degree of sensitivity and limit the ability to resolve structures or to accurately perform more complex image processing techniques such as deconvolution. While, image compression does not impact the utility of images for documentation, and is generally required when submitting images for publication, compression essentially discards information required for high-quality image quantitation.

It is also possible to inadvertently compress data by choosing a file format with inherent compression present. File formats can be broken down into two categories: “lossy” formats, which have built in compression algorithms, and “loss-less” formats, which maintain all pixel data. Formats like Joint Photographics Expert Group (JPEG) or Graphics Interchange Format (GIF) are considered lossy, or partially lossy, whereas formats like Tagged Image File Format (TIFF) are loss-less, and maintain the full bit depth and pixel values for all pixels. TIFF is the preferred, non-proprietary loss-less file format as it is a readily accessible by many current analysis software platforms, including open-source image-analysis software. While, the cross-platform utility of TIFF files retains flexibility, it lacks the file organization and convenience of proprietary file formats, which may or may not be as universally applicable. Proprietary formats can also help organize otherwise unwieldy datasets to streamline analysis, but are limited to a few software applications. Ultimately, the analysis software utilized by an investigator should dictate the file format. When considering these options, it is important to remember that documentation images have more file-format flexibility, whereas all data destined for quantitation should be stored as an uncompressed image file.

Three-Dimensional Imaging Methods

Imaging along the z-axis allows for accurate volumetric measurement, but also compounds the size of the image file. Increasing the number of z slices while optimizing z-step size enhances the ability to resolve structures in three-dimensional (3D) space, so long as Nyquist sampling is maintained. Practically speaking, the spacing between successive optical sections should not exceed 2.3× resolvable elements—keeping in mind that z-axis resolution is typically only half the resolution of the x and y axes, and is dependent upon the choice of objective and wavelength, as described above.

Data volumes in 3D are generally single-frame images collected as a depth projection. These image stacks are multi-dimensional, and consequently data handling and organization are somewhat more complex: each image field is composed of multiple individual files that may be stored in various proprietary formats, which conform to the microscope manufacturers. Forethought as to what will be done with the images should be considered when acquiring and saving these datasets, as analysis and viewing of such multi-dimensional files between platforms can be difficult. Somewhat fortunately for data handling, photo-toxicity or fluorophore bleaching often limits the volumes that can be collected with maximal signal range, resulting in smaller file sizes.

Four Dimensional Imaging Methods

A core principle of modern microscopy is the ability to address advanced biological questions over a volume in time. Such experiments can provide kinetic information regarding protein production, degradation, and localization; quantify ionic changes; define signal transduction pathways; and/or visualize cell proliferation and/or migration. A truly multi-dimensional, complex-systems approach must be taken to address such questions. Yet, adding the time component, as per Nyquist sampling limits, inevitably and dramatically increases the size of datasets. For example, a relatively modest four-color (blue, green, red, far-red and/or a brightfield channel), 3D (10 z slices), 50 time point, sCMOS image set will be 25 gigabytes—which is a trivially small dataset, in light of most four-dimensional imaging practices.

STORING FAST-IMAGING DATA

Data generated by a fast microscope commonly exceeds the writing-speed capabilities of a conventional hard drive or RAID (redundant array of independent disks) system. Implementing a solid-state-drive RAID 0 array—generally about 1–2 terabytes in size—is therefore, essential for systems dedicated to fast-imaging modalities. As described below, this type of storage must still remain volatile; in fact, users should be required to remove data from these systems immediately upon completion of an experiment, since the 1–2 terabyte capacity is still limiting—especially for complex, four-dimensional experiments.

THE PROBLEM OF DATA MANAGEMENT

The impact of modern microscopy on all aspects of biomedical research is rapidly expanding. A current trend is in integration of all the examples described above: Time-lapse, high-throughput/content screens performed in 3D, acquiring 32 channel spectral datasets, with a reliance on high-speed robotics that require faster systems to keep up with and handle the data. In another example, super-resolution approaches often demand 30,000 images to generate a final picture, in the case of stochastic optical reconstruction microscopy (STORM), or 15 images to generate a single plane of a one-channel, 3D structured-illumination microscopy (SIM) image.

Centers that specialize in microscopy rarely rely on a single microscope. A large multi-user facility may have multiple wide-field, confocal, live-cell, and super-resolution systems—all of which demand data bandwidth. The data volume generated is constantly expanding, yet there is no rational solution to storing and maintaining all of it on facility servers or systems.

Traditional Data Storage Issues in a Core Environment

Data are traditionally stored on a dedicated collection of local hard drives, with files stored on the machine on which they were generated. Analysis and manipulation is local and performed by the operator. While, this may work in the rare environment where each user has a personal microscope, this is completely impractical in a shared-resource environment where the instruments are in high demand and are under continual use.

Microscope-based acquisition systems are the core of any imaging facility, and are extremely expensive. Tying up an acquisition system computer for image processing or analysis is thus foolhardy, as it will inevitably limit microscope access and image acquisition time. Another related obstacle is that computers dedicated to running basic imaging systems with slow image-acquisition speeds are generally not that powerful; high-end image processing, particularly for 3D-over-time experiments, demands a system dedicated to image analysis—for example, those with fast processors and significant memory. The ideal solution is the co-existence of two dedicated platforms: one for imaging, and one for analysis. True, the data analysis software can be costly—but computers are relatively cheap, especially when considered in relation to the cost of a research microscope.

Networking: Where and When to Store

A central file server can drastically increase the ease of data writing and access for shared-facility users. Data can be collected on an acquisition system, and immediately transferred to the facility server, which keeps the imaging system free for another user to collect data. With such an offline system, users can access their collected data using any number of dedicated and networked analysis systems. One solution is the Dell Power Edge server consisting of 4-terabyte hard drives in a RAID format, although, most of the server options from notable hardware technology companies will perform adequately; building a similar server from individual components is also a reasonable solution, provided the on-site information technology group is willing and capable of maintaining it.

RE-IMAGINING THE PLATFORM FOR DATA-INTENSIVE TECHNIQUES

As described at great length in the above sections, data volume continues to grow. Imaging facilities typically use local storage systems that back up to tape or DVD, or rely on institutional services for the same function. This traditional approach, remains a plausible solution for images generated and stored from a standard light microscope or confocal microscope as these datasets are in the 10- to 100-megabyte range, which readily archive to a DVD. Fast confocal systems or microscopes using sCMOS cameras, however, generate data volumes that cannot be rationally archived in a rapid way, since they can produce enough data to fill a DVD every 2 sec; thus, a different approach is required. These realities have given rebirth to the “sneakernet”, in which data are physically moved among different sites via physical hard-drives as opposed to a network; in a contemporary guise, the user takes control of his/her data as much as possible. The following operational model describes a series of solutions that together supports 250–300 groups annually on 29 large stands, for a purchase price of over $250 thousand.

Networking

Large-scale imaging facilities generate massive amounts of data. Accordingly, data storage and distribution in facilities should: (i) Be minimally equipped, but utilize gigabit networking throughout. (ii) Exist behind a secure firewall to avoid extreme hacking into facility systems, and prevent access “into” the facility network from outside sources. (iv) Limit storage space within the facility because, regardless of the space available, it will quickly become full. In that guise, storage capacity should be limited to no more than two months of rolling use, since few users routinely access data older than 2 months (at this point in time, DVD archives are the best compromise for archiving and access; alternatively, archive onto hard drives rather than tape drives, since image datasets are big). (v) Have a dedicated file-transfer-protocol (FTP) or data-distribution site within the facility to allow users to post data for sharing or collection. (vi) Have a rational, appropriate, and consistent networking archive structure that is rigorously maintained and includes detailed information regarding the user group, the individual, date, and experiment so that searches are easily performed.

User Responsibilities

Dense, portable storage devices are increasingly small and fast, and are quite affordable. As such, individual researchers should be able to manage much of their own data. While a shared-use facility can and should archive large volume of small datasets (less than 100 megabytes), all users should be encouraged and empowered to manage their own larger datasets. General rules for handling the different sizes of datasets include: (i) All small datasets are immediately archived to online, fast storage (no local storage). (ii) All large datasets (>100 megabytes) must be managed by the user, and must be removed from the server, or microscope within a week. (iii) Users are expected to have access to physical, portable USB storage devices (either disk or RAM-based). (iv) While the facility will attempt to protect data, the data becomes volatile if/when space is needed. Under these conditions, all users are expected to maintain their storage devices in a manner in accordance with those of the institution’s digital security policies. Considering, the ability for device-embedded security threats to disable or slow facility systems, users should always take proper care to avoid inadvertently introducing malware or spyware to the shared-facility systems.

Archiving Responsibilities

Data space is always limiting, so users should have specific and realistic expectations of what the shared-resource facility can archive. For example, users should assume that: (i) No datasets will be archived by the facility; instead, all microscopes with high speed data capabilities will be equipped with USB3 ports for users to take their data away using personal storage devices. (ii) The facility will allocate a fixed volume of data storage per user (e.g., one DVD per user per month). (iii) Data should not be stored locally on microscopes whose data-acquisition rate is slower than network speed. (iv) Data stored locally on microscopes are considered volatile after one week. (v) When data are archived to DVD, they will be indexed and organized such that users can easily find the directories. An abundance of shareware that also validates the DVDs is available for this purpose (e.g., ThumbsPlus from Cerious Software).

Analysis Computing

Ensuring that the microscopes have maximal use for image acquisition, all analysis, and quantitation should be done on machines other than those controlling the microscopes. Given differences in needs between acquisition and image processing, each analysis, and quantitation computer should be equipped with adequate memory and processing power to accomplish any of the tasks a user requires of the software. Within a facility, the number of machines available for offline analysis should match the users’ needs (e.g., one analysis machine per three microscopes, loaded with appropriate software for each). Finally, computing resources should be available as close to 24/7 as possible due to the unpredictable processing times.

CONCLUSIONS

Microscopes are becoming faster and more capable, able to collect data in four-dimensions as well as spectrally, using a multi-well, or multi-field stitched format. Consequently, a wellrun, large shared-use imaging facility can readily generate petabytes of data each month, leading to extraordinary frustration with regards to how these data are managed. While, there is no definitive solution to the evolving problem of data storage, an imaging facility can ensure autonomy by:

- Becoming isolated from the backbone of its institutional or corporate storage solution.
- Expecting users to manage their own data.
- Providing local resources such that users can perform all analysis tasks in house.
- Providing local connectivity such that users can use local storage devices at any location.

These general rules, along with expert guidance by facility personnel on the utilization of the microscopy systems as well as computing systems, provide a streamlined framework for users to fully realize the advanced capabilities of the imaging systems and techniques available–in order to gather cutting-edge experimental results.

Abbreviations

3D	three dimensional
CCD	charge-coupled device
[s] CMOS	[scientific] complementary metal oxide sensor
RAID	redundant array of independent disks