1084
doi:10.1017/S1431927621004086
Microsc. Microanal. 27 (Suppl 1), 2021

Centralizing digital resources for data management, processing, and analysis for enterprise scale imaging research

Mathieu Gendron[1], Jean-Frédéric Fontaine[1], Benjamin Provencher[1], Eric Yen[1], Nicolas Piché[1] and Mike Marsh[2]

[1]Object Research Systems, United States, [2]Object Research Systems, Denver, Colorado, United States

If you knew what you were going to do before you got started, it wouldn't be research. That great uncertainty underlies the vast challenge of recording experimental parameters and results even as the experimental protocols themselves evolve over the duration of a research project. This problem of electronic lab notebooking is compounded further when the experiments involve images. The need to store, search, access, and interpret scientific images which are captured in a variety of file formats with highly variable--and often unanticipated--metadata quickly overwhelms established solutions for tracking, cataloging and interoperating with such data. While these issues are tractable for smaller studies, modest solutions fail to scale and become unmanageable when the volume of images and associated parameters are acquired at the enterprise scale. The ultimate decisions on infrastructure need to support the FAIR principles of findability, accessibility, interoperability and reuse of digital assets [1]. We describe here how a centralized framework of data traceability, data organization, visualization, and compute resources provides key interoperability among instruments and software for research imaging. We carried out this work with the commercial platform Dragonfly for data organization and visualization, but the strategic scaffolding of resources should serve as a model that can be replicated by other software ecosystems.

Before we organize the data and compute resources, we consider the implications of proper record-keeping. To support reproducibility and data mining of research methods, it is paramount that the image operations and history are recorded properly in the framework. Research accountability demands traceability of metadata on individual images and upstream parameters (e.g. sample preparation), as well as downstream parameters (e.g. image processing parameters, data mining parameters, etc.). Our image processing expertise drew us to focus first on metadata capture and encoding in that scope. We have implemented a framework of logging that tracks and records relevant metadata [2]. This means that users might experiment normally with different image processing options and parameters and workflows, and that when they save the resulting image or analysis, the log of operations is encoded in the file. The user can extract the process log as a fully functional processing macro to reproduce the exact work on that file, or to batch process the same protocol on the broader experimental dataset. This means that documenting the workflow is not vulnerable to poor note-taking. It means that automation of complicated workflows is immediately available with no advanced planning or even any programming expertise. We believe this solution of appropriate logging is a model for encoding other metadata from image acquisition, sample preparation, computer modeling, data analysis, etc. With proper logging, all experimentation becomes self-documented and ready for up-scaled automation.

The core challenge is to organize the data and metadata so that nothing is ever lost such that every detail can be found by browsing or searching--through both human interaction or software-directed queries from compute resources. The central data store must serve as a hub, linking image production and image consumption and research findings. By image production, we mean raw image acquisition and the production of derivative images produced by image filtering, image segmentation, etc.. By image consumption, we mean everything from the interactive inspection of images on to image-informed

computational modeling. We use a lightweight project organizer where users can deposit and maintain any kind of file, linked with the rest of the data for that project. Users can ascribe search tags to individual projects or associated files, making it easy for workers to navigate and search vast libraries of projects. This project organizer serves findability for interactive use, but it also serves as the hub to connect the data to the compute and visualization resources

Support for interactive visualization and computation as well as asynchronous processing for more compute-intensive tasks on distributed compute resources must be handled properly. Some solutions may give superficial 2D or 3D visualization from remote resources, but then rely on the user to download the data before performing computation and analysis; we aim to avoid this download of data altogether. It is our observation that when processing is distributed among individual workers' desktops, data and work history records are mismanaged and often lost. We keep data in the datacenter by establishing a fully connected 3D visualization and distributed computing and bookkeeping solution. Our visualization solution uses remote rendering to meet the dual mandate of never putting data at risk while still providing users the full suite of normal interactive tools from any remote connected laptop, workstation, tablet, etc. All data are tagged with unique identifiers which means that any compute task must only know its inputs and parameters, before it can proceed to operate, compute results, and then deposit results back into the appropriate project. We also provide a batching system that lets users submit compute jobs to a processing queue. The compute jobs themselves are fully aware of their respective data store origins and destinations. The consequence is that batch jobs can proceed asynchronously, and the user will discover the results in the expected project space when the task is complete; of course the user can monitor the progress of the queue to see the expected completion time. We make the task queue open and extensible to any sort of data mining, numerical modeling, or other compute-intensive task; this openness and extensibility to third-party development permits the ecosystem to grow beyond the current vision and scope.

References

[1] Wilkinson, M. *et al. Sci Data* **3,** 160018 (2016).
[2] Robert, F. *et al. Microscopy and Microanalysis,* **26**(S2), 1714-1715 (2020)

https://doi.org/10.1017/S1431927621004086 Published online by Cambridge University Press