CICompass

TECHNICAL REPORT

ci-compass.org

The Major Facilities Data Lifecycle in a Nutshell

Laura Christopherson, October 5, 2021

The CI CoE Pilot project was funded in 2018 and charged with creating a blueprint for a cyberinfrastructure (CI) Center of Excellence, dedicated to supporting NSF major facilities' (MFs) CI that is critical to achieve science outcomes. To better inform the Pilot's creation of that blueprint, we began researching the way MFs manage their data and the cyberinfrastructure (CI) that enables the capture, transformation, and dissemination of science data. We selected a small set of MFs for the initial analysis: IceCube, the Laser Interferometer Gravitational Wave Observatory (LIGO), the Rubin Observatory, the National Ecological Observatory Network (NEON), and the Oceans Observatories Initiative (OOI)[1]—based on the following criteria:

- Together, these MFs employed a broad range of CI tools, services, and architectures.
- We presumed that, across all of them, they may face a diverse set of challenges when managing their data.
- They each occupied a different phase of an MFs' lifecycle (e.g., planning and construction, operations, and facility maturity involving enhancements and upgrades to the CI).
- We had existing relationships with these MFs, which would make interviews and fact-checking straightforward.

We began by conducting a scholarly literature review of papers written by MF staff and examining publicly-available design documents, specifications, progress reports, policies and procedures documents, and MF websites. Then we interviewed staff at these facilities to confirm what we had learned and to seek answers to any open questions. From this review, common themes emerged across these different facilities, which crystallized into a general model that described how science was done at large, earth science research facilities. Although very different in many ways (e.g., different units of analysis, different equipment used to capture data, different forms of data processing, etc.), we learned that MFs tend to follow the same general path from the moment of data capture to the moment of data dissemination.

We then conducted a literature review to determine if any existing work investigating data/research lifecycles adequately captured the data flows we observed with NSF MFs. There has been considerable work done to summarize and model data/research lifecycles across a variety of disciplines (see [1] for more information.) However, none of the pre-existing models we studied completely paralleled the NSF MF data lifecycle, and so would not have been a sufficient guide for understanding NSF MFs better. As a result we created a new lifecycle model to capture the way data flows within the MFs and to guide us in supporting their work and developing the blueprint (see [2] for more information) for a center to continue that support.

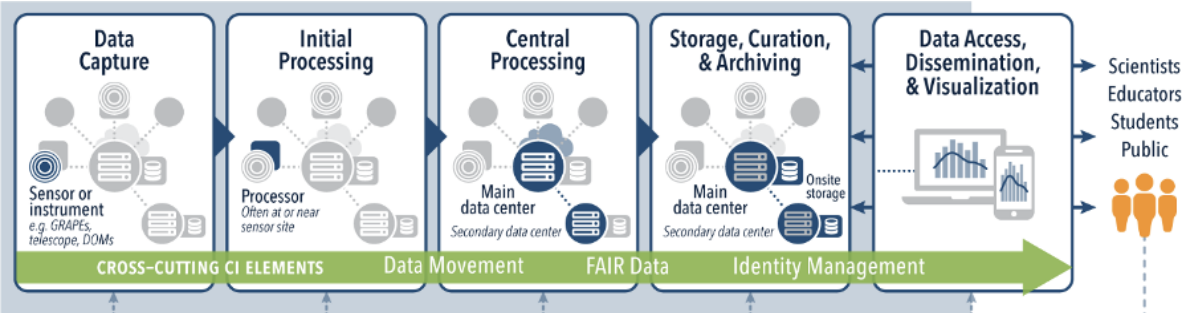
Data Lifecycle Stages Described

The stages of this model are described as follows:

Data Capture — As can be imagined, all the MFs we reviewed perform some sort of data capture at their instrumental facilities. For example, LIGO captures wave forms from its two interferometers. Rubin intends to capture

[1] IceCube (https://icecube.wisc.edu), LIGO (https://www.ligo.caltech.edu), Rubin (https://www.lsst.org), NEON (https://www.neonscience.org), OOI (https://oceanobservatories.org)

This project is supported by the National Science Foundation Office of Advanced Cyberinfrastructure in the Directorate for Computer Information Science under Grant #2127548.



The NSF Major Facilities Data Lifecycle

images from its telescope. NEON captures data from field sensors, tablets, used by scientists in the field, and remote sensing airplanes that fly over field sites. OOI captures data from sensors on cables on the ocean floor and buoys on the ocean's surface.

Initial Processing — Most of the MFs we examined perform initial filtering and processing. Usually, this is conducted at the capture site or nearby, and is intended to prepare the data for later transmission to a data center for more involved processing and analysis. Initial processing may also be conducted to alert the MF to particularly interesting scientific events that require immediate attention. For example, IceCube generates alerts and reduces the volume of the data at the South Pole, making it ready for faster transmission to its data center in Wisconsin. Similarly, Rubin intends to generate real-time alerts as data is captured and will prepare the data for its offline analysis by performing detector cross-talk correction and creating metadata. LIGO initially down-samples their data from 16k Hz to 4k Hz, which not only reduces data volume making it more manageable, but also eliminates considerable noise from the data.

Central Processing — Central processing may involve additional cleaning and data preparation, quality control measures, and/or the application of algorithms and transformations to make the data science-ready. Currently, NEON has one data center in Denver, Colorado, and performs the bulk of its data processing there. The data is calibrated, physical units are converted into standard scientific units, quality control measures are applied, and gaps in time in the data (due to collection at multiple sensors) are resolved. In addition, different levels of data products are systematically generated using several scientific transformations that leverage automated processing infrastructure in their data center. OOI conducts the bulk of its processing at its data center (at Oregon State U.). This processing involves various quality control measures, creating calibration information, formatting the data for later analyses, generating metadata, and performing specialized processing upon user request. LIGO may aggregate neutrino events into "superevents" if they are close in time and apply Monte Carlo simulations to compare different search techniques. In Wisconsin, IceCube conducts the bulk of its processing on filtered data sent from the South Pole, and then uses distributed resources (e.g., Open Science Grid, XSEDE resources, campus clusters, NSERC[2]) to generate further levels of science-ready data.

Storage, Curation, and Archiving — MFs archive and store data for the purposes of retaining a history of observations across time and ensuring its availability to and use by affiliated scientists and the general public. Some MFs replicate data from the main data center to other locales, such as NEON, which keeps replicas of its Denver data center data in its Boulder headquarters and on the cloud. Some store data in a variety of locations on different forms of media. For example, IceCube stores copies of data at the South Pole, Wisconsin, California, and Germany. Some data is stored on disk, some on tape. The Archiving/Storage stage is critical because the data is a record of the facility's fulfillment of its science mission.

Data Access, Dissemination, & Visualization — All MFs are required by NSF to disseminate their data. Usually, they distribute the data first to

[2] National Energy Research Scientific Computing Center (https://www.nersc.gov/)

This project is supported by the National Science Foundation Office of Advanced Cyberinfrastructure in the Directorate for Computer Information Science under Grant #2127548.

their own collaborative or consortium of scientists, and later to the general public. Some of an MF's data, however, may be for collaborative/consortium eyes only. MFs tend to provide multiple avenues of data access, including download via web-based data portal or FTP (e.g., NEON, OOI, LIGO, Rubin), API/web services (e.g., NEON, OOI, Rubin), shipments of data (e.g., NEON), distributed data management systems (e.g., IceCube), or even an in-person visit to a data center (e.g., Rubin).

In some cases, the DLC stages can have dependencies among them, and they often have customized sub-stages, data flows, and loops that are nuanced with respect to particular MFs. Sometimes, data generated by users external to the MFs can flow back to various points in the DLC (e.g., analyzed data is used for steering instruments at the capture stage, higher level data products are added for dissemination). Some CI aspects are cross-cutting through DLC stages such as: 1) data movement functions, technologies, and policies; 2) data representation, ontologies, and cross-domain data discovery (FAIR data); and 3) identity management for data providers, administrators, and users, which is important to the safekeeping and policy-based sharing of the data.

Movement — For MFs, the transition from one stage to the next always involve moving the data from one instrument or location to another. If the data is not being captured, processed, stored, or disseminated, it is on its way somewhere. Hence, movement is a cross-cutting element of our data lifecycle model. Methods of movement include satellite (e.g., IceCube), undersea cables (e.g., OOI), conventional networks (e.g., NEON, Rubin), and physical transfer of hard disks by plane or other transportation (e.g., NEON, IceCube).

Application of the Data Lifecycle to Other MFs

Since these initial investigations of IceCube, LIGO, the Rubin Observatory, NEON, and OOI were completed, the Pilot conducted in-depth interviews with these other MFs to determine if the Data Lifecycle model shown above continues to resonate with other MFs.

- Cornell High Energy Synchrotron Source (CHESS) https://www.chess.cornell.edu/
- Incorporated Research Institutions for Seismology (IRIS): Seismological Facilities for the Advancement of Geoscience (SAGE): https://www.iris.edu/hq/
- Large Hadron Collider https://home.cern/science/accelerators/large-hadron-collider
- National Hazards Engineering Research Infrastructure (NHERI): Design Safe https://www.designsafe-ci.org/
- National Hazards Engineering Research Infrastructure: Reconnaissance Facility (RAPID) https://rapid.designsafe-ci.org/
- National Optical-Infrared Astronomy Research Laboratory (NOIRLab) https://noirlab.edu/
- University NAVSTAR Consortium (UNAVCO): Geodetic Facility for the Advancement of Geoscience (GAGE): https://www.unavco.org/

The staff interviewed from these other MFs concur that the Data Lifecycle model shown above does a good job of illustrating the way data is managed and flows through their facility.

References

1. Laura Christopherson, Anirban Mandal, Erik Scott, and Ilya Baldin. 2020. Toward a Data Lifecycle Model for NSF Large Facilities. In *Practice and Experience in Advanced Research Computing (PEARC '20).* Association for Computing Machinery, New York, NY, USA, 168–175. DOI:https://doi.org/10.1145/3311790.3396636
2. Ewa Deelman, Anirban Mandal, Angela P Murillo, Jarek Nabrzyski, Valerio Pascucci, Robert Ricci, Ilya Baldin, Susan Sons, Laura Christopherson, Charles Vardeman, Rafael, Ferreira da Silva, Jane Wyngaard, Steve, Petruzza, Mats Rynge, Karan Vahi, Wendy Whitcup, Josh Drake, and Erik Scott. 2021. *Blueprint: Cyberinfrastructure Center of Excellence*. https://doi.org/10.5281/zenodo.4587866

Note that much of the content that appears in this report was taken from [1].

This project is supported by the National Science Foundation Office of Advanced Cyberinfrastructure in the Directorate for Computer Information Science under Grant #2127548.