Toward a Data Lifecycle Model for NSF Large Facilities

Laura Christopherson
laura@renci.org
RENCI, University of North Carolina at Chapel Hill

Anirban Mandal
anirban@renci.org
RENCI, University of North Carolina at Chapel Hill

Erik Scott
escott@renci.org
RENCI, University of North Carolina at Chapel Hill

Ilya Baldin
ibaldin@renci.org
RENCI, University of North Carolina at Chapel Hill

ABSTRACT

National Science Foundation large facilities conduct large-scale physical and natural science research. They include telescopes that survey the entire sky, gravitational wave detectors that look deep into our universe's past, sensor-driven field sites that collect a range of biological and environmental data, and more. The Cyberinfrastructure Center for Excellence (CICoE) pilot project aims to develop a model for a center that facilitates community building, fosters knowledge sharing, and applies best practices in consulting with large facilities with regard to their cyberinfrastructure. To accomplish this goal, the pilot began an in-depth study of how large facilities manage their data during the course of their research. Large facilities are diverse and highly complex, from the types of data they capture, to the types of equipment they use, to the types of data processing and analysis they conduct, to their policies on data sharing and use. Because of this complexity, the pilot needed to find a single lens through which it could frame its growing understanding of large facilities and identify areas where it could best serve large facilities. As a result of the pilot's research into large facilities, common themes have emerged which have enabled the creation of a data lifecycle model that successfully captures the data management practices of large facilities. This model has enabled the pilot to organize its thinking about large facilities, and frame its support and consultation efforts around the cyberinfrastructure used during lifecycle stages. This paper describes the model and discusses how it was applied to disaster recovery planning for a representative large facility—IceCube.

CCS CONCEPTS

• **Information systems** → **Data management systems**; • **Applied computing** → *Physical sciences and engineering*.

KEYWORDS

data lifecycle, data management, cyberinfrastructure, large facilities, research computing, disaster recovery

ACM Reference Format:
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), July 26–30, 2020, Portland, OR, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3311790.3396636

1 INTRODUCTION

The pilot project—the Cyberinfrastructure Center of Excellence (CICoE) pilot [9]—is a National Science Foundation (NSF)-funded effort to develop a model for a center that facilitates community building, fosters knowledge sharing, and applies best practices in consulting with large facilities with regard to their cyberinfrastructure. Large facilities are "shared-use infrastructure, instrumentation and equipment that are … intended to serve the science community" [19, p. 1.1-1]. Large facilities serve physical and natural science disciplines such as astronomy, physics, geoscience, and biology.

These facilities manage scientific data collection equipment such as large telescopes, interferometers, and distributed sensor arrays. They collect terabytes of data every year and make it available to scientists for their research. Examples of large facilities include the IceCube Neutrino Observatory,[1] the Large Synoptic Survey Telescope (LSST),[2] the National Ecological Observatory Network (NEON),[3] the Ocean Observatories Initiative (OOI),[4] and the Laser Interferometer Gravitational-Wave Observatory (LIGO).[5] These facilities are responsible for a diverse set of equipment in multiple locations (sometimes spanning the globe). For example, NEON has approximately 81 field sites (terrestrial and aquatic) across the U.S. from which data is collected in three main ways: via airplane, field tablets, and streaming sensors. It processes the data in a data center located in Denver, Colorado. IceCube's neutrino detector is located at the South Pole, and its main data center is located at the University of Madison-Wisconsin. It stores data at the National Energy Research Scientific Computing Center (NERSC)[6] in California and performs further processing at the Deutsches Elektronen-Synchroton research center (DESY)[7] in Germany, among other locations.

Large facilities collect a variety of data types: images, video, sound, numeric measurements, human observations, physical samples, etc. They shepherd their data from its collection or creation

PEARC '20, July 26–30, 2020, Portland, OR, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-6689-2/20/07…$15.00
https://doi.org/10.1145/3311790.3396636

[1] IceCube, https://icecube.wisc.edu/
[2] LSST, https://www.lsst.org/
[3] NEON, https://www.neonscience.org/
[4] OOI, https://oceanobservatories.org/
[5] LIGO, https://www.ligo.caltech.edu/
[6] NERSC, https://www.nersc.gov/
[7] DESY, http://www.desy.de/

to its filtering and cleaning to its processing and analysis to its storage and ultimate dissemination. Because they tend to oversee data as it travels from collection instruments (e.g., sensors) to computers and clusters to consumable media (e.g., website files, disks), crossing multiple geographic boundaries, large facilities are faced with a variety of challenges and complexities in managing and caring for their data. The CICoE pilot's job is to envision how to best aid large facilities in better managing this intricate process while maximizing their use of the end-to-end cyberinfrastructure. This includes considering how to best support them in matters of data capture, processing, visualization, security and identity management, storage and preservation, dissemination, and disaster recovery. To accomplish this, we needed to first understand the way large facilities use their cyberinfrastructure to conduct their day-to-day operations and how they meet research and data challenges as they arise.

We began by surveying a small set of large facilities to identify their data management practices and the cyberinfrastructure they use. The process of learning their strategies has involved a combination of content analysis of publicly available documents (e.g., scholarly publications, large facility websites, large facility document archives), inference based on learned facts, and clarification and confirmation of details via interviews and engagement with facility staff.

Each large facility is a unique, highly complex organization. This complexity is amplified when considering multiple facilities. So we needed to find a single lens through which we could frame our growing understanding of large facilities and better organize our thinking about them. Regardless of the differences in the type of science studied, the different types of data created/used, or the variation in how research operations are conducted, we found that the large facilities we reviewed tend to traverse a similar path when caring for data throughout its lifecycle. This shared data lifecycle is the lens through which we have begun to frame our work with large facilities; and the purpose of this paper is to report on our efforts to develop that lens, or if you will, the data lifecycle model.

The main contributions of this work include:

- the introduction of a data lifecycle model, that, to our knowledge, has not been previously created to describe end-to-end data operations for NSF large facilities,
- a lens through which to better compare and contrast the ways different large facilities conduct research operations and manage data, and
- the application of this model to an important aspect of large facility cyberinfrastructure—disaster recovery planning; specifically, the model was used to provide context and structure a template used for disaster recovery planning.

In this paper, we will report on other lifecycle models and discuss why a data lifecycle model for large facilities is needed. We will then present and describe the data lifecycle model, and show how we have applied it to the task of disaster recovery planning for a representative large facility—IceCube. Then we will identify future steps needed to develop this work further. We anticipate this model will continue to evolve as we expand our research to other large facilities.

2 LIFECYCLE MODELS

Data and research lifecycle models abound in the literature. Together [3, 5–7, 26, 30] examine well over 50 models. These models have many attributes in common such as shared stages, but many differences as well, particularly with regard to purpose, intended audience, actors described, and focus. We reviewed these models to learn from the work of their creators. Specifically, we wanted to identify any commonalities that could help us validate aspects of our own model and any different qualities that we should include.

Data lifecycle models are created for a variety of reasons, such as to promote good data curation and preservation practices (e.g., [1, 15]), to encourage researchers to manage their data more effectively and make it more available for reuse (e.g., [2, 13, 14, 25, 29, 31]), to improve data management in specific scenarios such as those that involve big data or employ cloud resources (e.g., [10, 18, 21, 28]), to expose available library services or resources needed at different points in the research process (e.g., [24]).

Models are directed toward different audiences, such as researchers, data managers of various kinds (e.g., preservationists/archivists, librarians, repository administrators), systems and software developers, and data providers. Some models are created to describe and/or aid very specific audiences with specific interests, such as those interested in digitizing content or creating digital content [11], users of linked data [4], cloud customers [21], those that need to see evidence of the value of open research [23], and those who already use or would need to use a particular standard [17, 20]. Models may be intended to guide the individual or an entire community or organization. Furthermore, models are designed to describe and guide practitioners in a variety of domains, including industry and business [10, 18], physical sciences [1, 2, 13, 27, 29, 31], natural sciences [2, 23, 31], social sciences [14, 17, 27], and technology [4, 21, 23].

Models are also developed with a particular focus, usually emphasizing either the larger research process, technical aspects of the process, or data-oriented aspects. For example, models that focus on the larger research process may have stages for things like study design or proposal writing (e.g., [16, 17, 23]). Those that emphasize technical aspects may include information about equipment or software used (e.g., [4, 18]). Those focused on data may include information about things like provenance, ethics around data use (e.g., confidentiality), particular types of data cleaning and preparation tasks, or even particular types of analyses that can be performed on the data (e.g., [14, 17, 18]).

Most model creators do not specifically describe how they derived their model (or, at least we found little detail about their process). In these cases, we assume that they formed a conceptualization of the research/data process through discussion and debate among themselves, and that this conceptualization was based on their close experience with the subject matter and the actors involved. When methods are elucidated however, they tend to include the review of scholarly literature, examination of artifacts used or produced in the process of research, and/or observation and interviewing of actors (e.g., [8, 23, 27]). Some models are adapted from other models and improved upon through more conceptualization or research to account for new situations and needs.

Most models are intended to represent the ideal situation [6] for the purpose of acting as a guide for establishing better habits and processes. Most models are high level representations, with details omitted, and they may attempt to cover multiple situations [6]. Carlson [6] argues that this obscures the richness and depth of these situations and fails to adequately delineate the differences between situations. While this might be true, the differences and details can be so varied and complex that we would argue it is impossible to delve too deeply into specific nuances or unique aspects of the situation without overcomplicating the model and potentially confusing the audience. For example, the data lifecycle and research process can be extremely varied with respect to the types of data used (e.g., text, audio, images), data collection methods employed (e.g., interviews, surveys, sensors, observation, recordings), the quantity of data collected, the data preparation and cleaning methods used (e.g., manual text cleaning, use of scripts to preprocess text or filter/reduce the size of streaming data), the types of analyses typically used which often vary by discipline, and any requirements imposed by funding agencies on the storage, protection, and dissemination of data. Although it is important for the model to be sensitive to variation, like details must necessarily be grouped into overarching themes. Only truly distinct and compelling characteristics should be highlighted if the model is to be understandable and useful to the widest possible intended audience. Carlson also contends that models are necessarily biased—reflecting the views and interests of those who create them. Again, this may be true in many cases, but there are instances where model creators have attempted to survey and study a range of situations (e.g., [8, 23, 27]) which would hopefully yield a more impartial depiction. Not to mention in most of the cases we examined, models were being created to suit specific needs and uses which necessitates the use of a particular lens. For example, the model that the Inter-university Consortium for Political and Social Research (ICPSR) [14] proposes is "biased" in its inclusion of details specific to the ICPSR data repository (e.g., details about using a metadata standard adopted by ICPSR). However, were it not to include these aspects, the model would not be achieving one of its aims, which is to instruct individuals on storing their data in ICPSR's repository.

These models and ours share many similarities. For example, almost all have some sort of stage for capturing/collecting, ingesting, or creating data, as ours does. However, there are sufficient differences in purpose, intended audience, actors described, focus, and the content of lifecycle stages that we concluded none would serve as an accurate depiction of large facilities. So we believed that a model specific to large facilities was needed. To our knowledge, such a model has not been created until now. **The purpose of our model is to guide our research into and engagement with large facilities, specifically at the intersection of data management practices and cyberinfrastructure.** *It represents an actual situation rather than any set of ideal or recommended practices. It is a high level representation of an outsider's perspective on the situation, but designed intentionally to describe a specific community (i.e., large facilities) and to satisfy a specific need (i.e., to support the pilot's mission to aid large facilities with using cyberinfrastructure for research).*

3 OUR APPROACH

To better understand NSF large facilities, we began by researching these five large facilities: IceCube, LIGO, LSST, NEON, and OOI. We selected these five because:

- A broad range of cyberinfrastructure tools, services, and architectures are employed across these five facilities.
- We believed these five would experience sufficient variation in the types and magnitudes of challenges faced when managing data.
- They offered the opportunity to view large facility needs and activities at different phases in a facility's lifespan (i.e., LSST is under construction; IceCube, NEON, and OOI are in operations; and LIGO has matured to the point of making enhancements in their operations and equipment).
- We have existing familiarity and long-term experience working with some of these large facilities, and so establishing contact for interviews and fact-checking would be straightforward.

For each facility, we selected material to review based on what was publicly available and relevant to our focus—data management practices and cyberinfrastructure used. Our information gathering methods included:

- scholarly literature review,
- examination of design documents and preliminary specifications created during the process of planning and construction of large facilities,
- examination of progress reports, revised/updated design documents, and finalized specifications,
- review of large facilities' websites, and
- interviews with large facility staff.

So for example, we reviewed journal papers written by IceCube developers about the data acquisition system they built, and collected design documents and images from LSST's document repository. Some of these facilities had multiple websites, so we reviewed them all. For example, LIGO has a site for the consortium/administrative body that oversees the facility (e.g., the LIGO Scientific Collaboration),[8] sites for different labs or facility locations (e.g., LIGO Caltech, MIT, Hanford, Livingston),[9] and a data portal (e.g., LIGO's Gravitational Wave Open Science Center).[10]

Detailed notes were taken on what was learned about each facility and then, as common themes emerged, facets of the model were derived per inductive content analysis [12, 22, 33]. For example, it became clear that all five of these large facilities capture data in a fairly involved manner that may vary based the nature of data or the equipment used to capture it. Data undergoes an initial processing phase, usually at the capture location, that may include filtering noise, down-sizing, adding calibration information, and/or identifying particularly interesting events that should be addressed immediately. As a result, our model needed to have a capture stage as well as an initial processing stage separate from the more in-depth processing that takes place sometime after capture, usually at the facility's data center.

[8] LIGO Collaboration website, https://www.ligo.org/
[9] LIGO facility website, https://www.ligo.caltech.edu/
[10] GWOSC, https://www.gw-openscience.org/

Carlson [6] suggests that when creating lifecycle models, the following be considered and factored into the model's design: scope, the best practices of other model developers, the importance of representing real-world activities, and the need to give attention to the transition between phases.

Scope — Our scope can be defined as follows:

- Actors described: Large facilities *(e.g., We did not observe characteristics of research institutes or labs at universities).*
- Topic/focus: Data management practices within the context of cyberinfrastructure use *(e.g., We did not review information about how research projects were managed or how the budget was used to purchase equipment).*
- Intended audience: CICoE team members, the broader NSF cyberinfrastructure community, large facility cyberinfrastructure managers and architects, and others interested in the work of large facilities.
- Purpose: To help the CICoE team frame and categorize their research into large facility cyberinfrastructure, in order to capture and develop best practices.

Our model was also created with the intention to *balance the level of detail needed to summarize commonalities while also highlighting the most compelling nuances.*

Best practices — As stated earlier in this paper, we reviewed the work of others who have created lifecycle models and used the lessons learned from these efforts to inform the design of our model. For example, because models that had the cleanest and most simplified representations were the easiest to understand, we mimicked this simplicity in our own model. Many models were composed mainly of boxes and arrows, but other models included specific imagery that, in juxtaposition to the box-and-arrow backdrop, drew attention to specific aspects of the process. We adopted this for our Dissemination stage, to draw attention to the user community.

Real-world activities — Our entire effort was grounded in understanding the real-world activities of large facilities. We reviewed factual documents that described the design, development, and use of large facility cyberinfrastructure for research and data management. We interviewed large facility staff to verify what we had learned and gather their feedback on our understanding.

4 THE MODEL

Our data lifecycle model, shown in Figure 1, represents our current understanding of large facility operations and how data is shepherded through the research process. Our goal in creating the model was to identify the overall commonalities across large facilities with respect to the services offered, functions performed, and cyberinfrastructure used at each lifecycle stage. It is a high level model, not intended to capture all the different nuances of various large facility operations. As such, it consists of five main stages: Capture, Initial Processing, Central Processing, Archiving/Storage, and Dissemination.

Capture — As can be imagined, all the large facilities we reviewed perform some sort of data capture using some sort of instrument. For example, LIGO captures wave forms from its two interferometers. LSST intends to capture images from its telescope. NEON captures data from field sensors, tablets used scientists in the field, and remote sensing airplanes flying over field sites. OOI captures data from sensors adhered to cables on the ocean floor and buoys on the ocean's surface.

Initial Processing — Most of the large facilities we examined perform some sort of initial filtering and processing. Usually, this is conducted at the capture site or close by, 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 large facility 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, LSST intends to generate real-time alerts as data is captured, and will prepare the data for its later 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 preparation techniques, quality control measures, and/or the application of algorithms and transformations that make the data more 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 Rutgers data center. 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, NERSC) to generate further levels of science-ready data.

Archiving/Storage — Large facilities 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 large facilities 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, in Wisconsin and California, and in 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.

Dissemination — All large facilities are required by NSF to disseminate their data. Usually they distribute the data first to their own collaborative or consortium of scientists, and later to the general public. Some of a large facility's data, however, may be for collaborative/consortium eyes only. Large facilities tend to provide multiple avenues of data access, including download via web-based

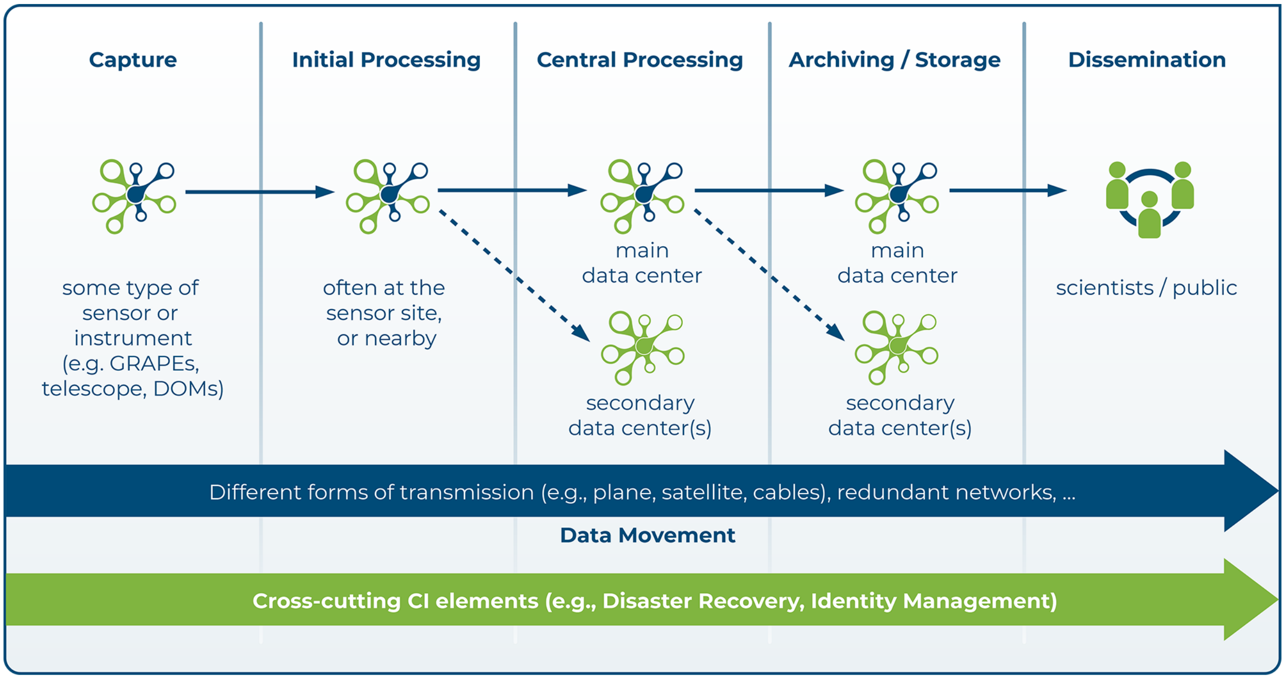

Capture
Initial Processing
Central Processing
Archiving / Storage
Dissemination
some type of sensor or instrument (e.g. GRAPEs, telescope, DOMs)
often at the sensor site, or nearby
main data center
secondary data center(s)
main data center
secondary data center(s)
scientists / public
Different forms of transmission (e.g., plane, satellite, cables), redundant networks, ...
Data Movement
Cross-cutting CI elements (e.g., Disaster Recovery, Identity Management)

Figure 1: The data lifecycle model for large facilities.

data portal or ftp (e.g., NEON, OOI, LIGO, LSST), API/web services (e.g., NEON, OOI, LSST), 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., LSST).

Movement — Recall that Carlson [6] suggested that model developers pay particular attention to the transitions between stages. For large facilities, transitions 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, LSST), and physical transfer of hard disks by plane or other transportation (e.g., NEON, IceCube).

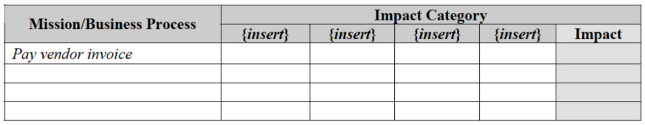
Other Dimensions — Our research into large facilities continues to uncover other possible dimensions that may need to be incorporated into the lifecycle model. Usually these are some form of attribute of the large facility that affects one or more stages, or data movement. For example, large facilities often evolve their data from its original state (which they often refer to as “raw") to some form of science-ready state. As the data is transformed from raw, to level 1, to level 2, and on upwards, it may require different handling. For example, large facilities may disseminate all of their data or just certain levels of their data; archival policies may vary based on the level of the data. Other dimensions we have been noticing include the types of research the large facility conducts (e.g., experimental, observational, computational), or whether the large facility is mobile (e.g., OOI which uses ships to rove the ocean) or largely stationary (e.g., IceCube). Because our research so far has not shown these possible dimensions to be applicable in all stages across all large facilities, we have not yet added them to the lifecycle model. We are still considering their influence.

Because it turns up repeatedly in our work, there is one dimension we suspect will mostly likely be added to the lifecycle—whether the facility captures, processes, archives/stores, and disseminates data in a distributed or centralized manner. For example, both NEON and OOI collect data from multiple locations, whereas IceCube and LSST collect from one location. NEON does most of its processing in one location, while IceCube distributes the processing between the South Pole, Wisconsin, California, and other locations. When the work is conducted in a distributed fashion, there may be more equipment to manage, more network lines to cross, and more potential points of failure. In other words, the cyberinfrastructure used may become more complex and incur more risk. As we move forward, expanding our work to other large facilities, we intend to give more thought to these other possible dimensions and revise the lifecycle model accordingly.

5 APPLICATION OF THE MODEL TO DISASTER RECOVERY

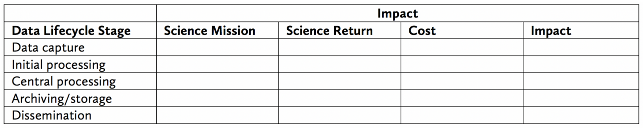
Part of fulfilling the pilot's mission involves advising and consulting with large facilities on topics such as identity management, data preservation, and disaster recovery. We first applied our data lifecycle model to the task of disaster recovery planning and selected NEON and IceCube as representative large facilities.

We began by creating a disaster recovery planning template based on the National Institute of Standards and Technology's (NIST) Contingency Planning Guide for Federal Information Systems [32]. Then we completed it for IceCube and NEON, and met with large facility personnel to discuss the template and make any



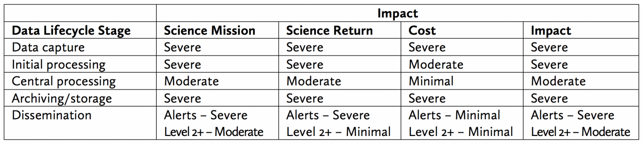
Mission/Business Process	Impact Category				
	{insert}	*{insert}*	*{insert}*	*{insert}*	Impact
Pay vendor invoice					

Figure 2: Sample impact matrix from the NIST Contingency Planning Guide for Federal Information Systems, p. B-3 (https://csrc.nist.gov/publications/detail/sp/800-34/rev-1/final).



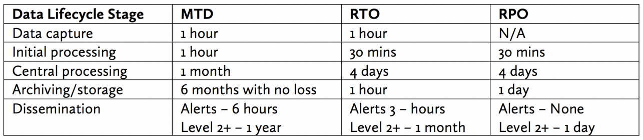
	Impact			
Data Lifecycle Stage	**Science Mission**	**Science Return**	**Cost**	**Impact**
Data capture				
Initial processing				
Central processing				
Archiving/storage				
Dissemination				

Figure 3: Impact matrix for large facilities adapted from the NIST Contingency Planning Guide for Federal Information Systems.



	Impact			
Data Lifecycle Stage	**Science Mission**	**Science Return**	**Cost**	**Impact**
Data capture	Severe	Severe	Severe	Severe
Initial processing	Severe	Severe	Moderate	Severe
Central processing	Moderate	Moderate	Minimal	Moderate
Archiving/storage	Severe	Severe	Severe	Severe
Dissemination	Alerts – Severe Level 2+ – Moderate	Alerts – Severe Level 2+ – Minimal	Alerts – Minimal Level 2+ – Minimal	Alerts – Severe Level 2+ – Moderate

Figure 4: Impact matrix for IceCube.



Data Lifecycle Stage	**MTD**	**RTO**	**RPO**
Data capture	1 hour	1 hour	N/A
Initial processing	1 hour	30 mins	30 mins
Central processing	1 month	4 days	4 days
Archiving/storage	6 months with no loss	1 hour	1 day
Dissemination	Alerts – 6 hours Level 2+ – 1 year	Alerts 3 – hours Level 2+ – 1 month	Alerts – None Level 2+ – 1 day

Figure 5: Estimated downtimes matrix for IceCube.

needed revisions. They were able to take the completed template and use it to formalize their own plans for disaster recovery. In this section, we will present disaster recovery planning examples from the IceCube template.

The NIST Guide asks organizations to conduct an initial impact analysis (i.e., take stock of their processes and their effects) to identify requirements needed for defining recovery procedures and resources. The impact analysis asks organizations to think about a variety of factors pertinent to disaster recovery, such as outage impact, estimated tolerable downtimes, and resources needed to respond to an outage or disruption. Organizations are encouraged to weigh these factors for each of their business processes or mission/goals. Large facilities share the same overall mission of conducting research, and, so far in our research, all operate along the same data lifecycle. So, for each of these mini-assessments, we customized the NIST template by using data lifecycle stages where organizations are asked to list specific business processes. The Guide urges users to revise "categories and values … to reflect what is appropriate for the organization" [32, p. B-2]. Applying the lifecycle in this way allowed us to offer a template that would be useful to all large facilities regardless of subtle differences in specific business processes.

Here we provide two examples of this customization—for the outage impact assessment and the estimation of downtimes. For the outage impact assessment, NIST suggests considering different ways a disaster could impact an organization, such as its cost, its potential to harm individuals, or its effect on the organization's ability to achieve its mission. These would be inserted as *Impact Categories* in the sample skeletal impact matrix shown in Figure 2. Specific business processes are to be listed in the first column.

For the impact matrix in our template, we adopted the *Cost* impact category NIST suggests, and created categories for *Science Mission* and *Science Return.* Stages of the data lifecycle are added to the *Mission/Business Process* column. The resulting impact matrix for large facilities is shown in Figure 3. Then we populated the impact matrix for IceCube with the appropriate severity values, shown in Figure 4, and confirmed our assessment with staff from IceCube.

Capture is an extremely critical function. If the in-ice sensors fail for more than a couple of hours, they will freeze solid, bringing them permanently offline, and so terminating all future facility operations. So they rated a problem with *Capture* as having *Severe* impact for each of the impact categories above. IceCube performs *Initial Processing* at the South Pole which involves generating alerts and performing some quality control measures (e.g., generating metadata and/or reducing data volume). Because of the urgency of alerts, IceCube rated their *Initial Processing* as *Severe* in all impact categories except *Cost*, which was rated as *Moderate*. Because a limited amount of hardware would be needed for replacing bad equipment, limited expense is incurred. Notice that in the *Dissemination* row, impacts are broken out by data type.

For our template's treatment of estimated tolerated downtimes, we retained the suggested column categories of *Maximum Tolerable Downtime* or *MTD* ("the total amount of time leaders/managers are willing to accept for a mission/business process outage or disruption and includes all impact considerations"), *Recovery Time Objective* or *RTO* ("the maximum amount of time that a system resource can remain unavailable before there is an unacceptable impact"), and the *Recovery Point Objective* or *RPO* ("the point in time, prior to a disruption or system outage, to which mission/business process data must be recovered") [32, p. B-3]. As with the impact matrix, stages in the data lifecycle were added to *Mission/Business Process* column. The result, along with cell values discussed with IceCube, is shown in Figure 5.

Because a failure in *Capture* would be *Severe* for IceCube, their estimated downtimes tended to be very short (i.e., in hours and not days). IceCube indicated 1 hour for *MTD* and *RTO. RPO*, however, was not applicable in this case because it is not possible to move back in time to the *Capture* stage in case of in-ice sensor failures.

Movement factors heavily in the estimated downtimes matrix. IceCube, for example, explained that *MTD* and *RTO* would vary based on where the data was in its journey from in-ice sensor, to the ice-top lab, to another South Pole lab before its transfer to Wisconsin. So far in our discussions with IceCube and NEON, we have addressed movement within the context of a given lifecycle stage. However, because movement pervades the entire lifecycle, it warrants its own consideration separate from individual stages. So we may need to revise our matrices to explicitly define it in rows of its own.

We anticipate the data lifecycle evolving as we learn more about other large facilities. As we consider the other dimensions mentioned in Section 4, we anticipate needing to change the template to reflect new knowledge. Even though our work is evolving and subject to revision, our efforts thus far have been beneficial for the pilot and the large facilities with which we have engaged. The data lifecycle has helped us frame our understanding of large facilities and apply it to disaster recovery planning. Customizing the NIST disaster planning guide by integrating the data lifecycle has enabled us to apply best practices to the cyberinfrastructure needs of large facilities. Engaging with IceCube and NEON around the disaster recovery template has furthered the pilot's goal of knowledge sharing and helped clarify the consultative role of an intended center more fully.

6 FUTURE WORK

Many more large facilities occupy our list for future research, such as the Academic Research Fleet,[11] Cornell's High Energy Synchrotron Source (CHESS),[12] the Large Hadron Collider,[13] and the National Hazards Engineering Research Infrastructure (NHERI).[14]

As we continue to learn more about other large facilities, we will necessarily revise the data lifecycle model accordingly. We don't anticipate having to add or delete stages, but if our research uncovers a need to do so, we will be open-minded to the fact. Currently, we anticipate needing to debate amongst our team the other possible dimensions that could be added to the model, dimensions that were mentioned in Section 4. These dimensions include the level of data, types of research conducted, mobile or stationary nature of the large facility, whether the large facility operates in a centralized or distributed manner, whether the large facility farms some of its compute or storage out or retains full control in-house. We anticipate that this list will grow as we continue our research into large facilities.

We also need to examine Movement more fully in regard to how it should be applied to the disaster recovery template. Currently it is not explicitly listed, separately from individual lifecycle stages. Up to this point, we have implicitly considered it when we have completed the template for a large facility. However, this is a tool we would like to give to large facilities to complete on their own, and they may not give Movement adequate attention if the template fails to explicitly prompt for it.

Future plans also include taking a survey of the hardware, software, and services that large facilities use in each stage of the lifecycle, and then making that information available to the entire large facilities community. This will enable large facilities to learn how their counterparts conduct operations, and if they would like to learn more about these other tools (e.g., pros and cons of use, how to configure them, lessons learned) they will know where they can go to learn more.

[11] Academic Research Fleet, https://www.unols.org/documents/academic-research-fleet

[12] CHESS, https://www.chess.cornell.edu/

[13] Large Hadron Collider, https://home.cern/science/accelerators/large-hadron-collider

[14] NHERI, https://hazards.colorado.edu/

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation grant OAC-1842042. We want to thank all members of the CICoE pilot project—especially the lead principal investigator, Ewa Deelman—who encouraged and supported the creation of the lifecycle model and our work on this paper. We thank Josh Terrell at RENCI for generating the figure for the lifecycle model. We'd also like to thank personnel from NEON and IceCube—in particular, Tom Gulbransen from NEON and Benedikt Riedel from IceCube. Both allowed us to learn from them about their data lifecycle, and offered valuable guidance and support to us in developing the disaster recovery planning template.

REFERENCES

[1] Sergio Albani and David Giaretta. 2009. Long term data and knowledge preservation to guarantee access and use of the Earth science archive. In *PV2018: Ensuring the Long-Term Preservation and Value Adding to Scientific and Technical Data*. 1–7.

[2] Suzie Allard. 2012. DataONE: Facilitating eScience through collaboration. *Journal of eScience Librarianship* 1, 1 (2012), 4–17.

[3] Mohammed El Arass, Iman Tikito, and Nissrine Souissi. 2017. Data lifecycles analysis: Towards intelligent cycle. In *2017 Intelligent Systems and Computer Vision (ISCV)*. IEEE, 1–8.

[4] Sören Auer, Lorenz Bühmann, Christian Dirschl, Orri Erling, Michael Hausenblas, Robert Isele, Jens Lehmann, Michael Martin, Pablo N. Mendes, and Bert Van Nuffelen. 2012. Managing the life-cycle of linked data with the LOD2 stack. In *International Semantic Web Conference*. Springer, 1–16.

[5] Alex Ball. 2012. *Review of data management lifecycle models*. University of Bath, IDMRC.

[6] Jake Carlson. 2014. The use of life cycle models in developing and supporting data services. *Research Data Management: Practical Strategies for Information Professionals* (2014), 63–86.

[7] Andrew Martin Cox and Winnie Wan Ting Tam. 2018. A critical analysis of lifecycle models of the research process and research data management. *Aslib Journal of Information Management* 70, 2 (2018), 142–157.

[8] Kevin Crowston and Jian Qin. 2011. A capability maturity model for scientific data management: Evidence from the literature. *Proceedings of the American Society for Information Science and Technology* 48, 1 (2011), 1–9.

[9] Ewa Deelman, Anirban Mandal, Valerio Pascucci, Susan Sons, Jane Wyngaard, Charles F. Vardeman II, Steve Petruzza, Ilya Baldin, Laura Christopherson, Ryan Mitchell, Loic Pottier, Mats Rynge, Erik Scott, Karan Vahi, Marina Kogank, Jasmine A Mann, Tom Gulbransen, Daniel Allen, David Barlow, Santiago Bonarrigo, Chris Clark, Leslie Goldman, Tristan Goulden, Phil Harvey, David Hulsander, Steve Jacob, Christine Laney, Ivan Lobo-Padilla, Jeremey Sampson, John Staarmann, and Steve Stone. 2019. Cyberinfrastructure Center of Excellence Pilot: Connecting Large Facilities Cyberinfrastructure. In *15th International Conference on eScience (eScience)* (San Diego, CA, USA). Funding Acknowledgments: NSF 1842042.

[10] Yuri Demchenko, Cees De Laat, and Peter Membrey. 2014. Defining architecture components of the Big Data Ecosystem. In *2014 International Conference on Collaboration Technologies and Systems (CTS)*. IEEE, 104–112.

[11] DigitalNZ.org. [n.d.]. *Getting Started with Digitisation*. https://digitalnz.org/make-it-digital/getting-started-with-digitisation

[12] Satu Elo and Helvi Kyngäs. 2008. The qualitative content analysis process. *Journal of Advanced Nursing* 62, 1 (2008), 107–115.

[13] John L. Faundeen, Thomas E. Burley, Jennifer A. Carlino, David L. Govoni, Heather S. Henkel, Sally L. Holl, Vivian B. Hutchison, Elizabeth Martín, Ellyn T. Montgomery, and Cassandra Ladino. 2013. *The United States geological survey science data lifecycle model*. Technical Report. US Geological Survey. https://pubs.usgs.gov/of/2013/1265/pdf/of2013-1265.pdf

[14] Inter-University Consortium for Political Social Research (ICPSR). 2012. Guide to Social Science Data Preparation and Archiving Best Practice Throughout the Data Life Cycle. https://www.icpsr.umich.edu/files/deposit/dataprep.pdf

[15] Sarah Higgins. 2008. The DCC curation lifecycle model. *International Journal of Digital Curation* 3, 1 (2008).

[16] Chuck Humphrey. 2006. e-Science and the Life Cycle of Research. https://era.library.ualberta.ca/items/3334684b-fa6a-4c9d-a74b-559fecd42f9f/view/79b064d6-7b51-4d18-8e4e-3d42b9faa81f/Lifecycle-science060308.pdf

[17] Data Documentation Initiative. 2019. Why Use DDI? https://ddialliance.org/training/why-use-ddi

[18] Nawsher Khan, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Zakira Inayat, Mahmoud Ali, Waleed Kamaleldin, Muhammad Alam, Muhammad Shiraz, and

Abdullah Gani. 2014. Big data: survey, technologies, opportunities, and challenges. *The Scientific World Journal* 2014 (2014).

[19] Finance Large Facilities Office in the Budget and Award Management Office (BFA-LFO). 2019. *Major Facilities Guide.* NSF 19-68. National Science Foundation. https://www.nsf.gov/pubs/2019/nsf19068/nsf19068.pdf

[20] Brian Lavoie. 2000. *Meeting the challenges of digital preservation: The OAIS reference model.* Technical Report. Online Computer Library Center (OCLC). https://www.oclc.org/research/publications/library/2000/lavoie-oais.html

[21] Li Lin, Tingting Liu, Jian Hu, and Jianbiao Zhang. 2014. A privacy-aware cloud service selection method toward data life-cycle. In *2014 20th IEEE International Conference on Parallel and Distributed Systems (ICPADS).* IEEE, 752–759.

[22] Philipp Mayring. 2004. Qualitative content analysis. *A Companion to Qualitative Research* 1 (2004), 159–176.

[23] Research Information Network and NESTA. 2010. Open to all? Case studies of openness in research. http://www.rin.ac.uk/system/files/attachments/NESTA-RIN_Open_Science_V01_0.pdf

[24] University of Central Florida Libraries: Scholarly Communication. [n.d.]. *Overview: Research Lifecycle.* https://library.ucf.edu/about/departments/scholarly-communication/overview-research-lifecycle/

[25] University of Virginia Library: Research Data Services and Sciences. [n.d.]. *Steps in the Data Life Cycle.* https://data.library.virginia.edu/data-management/lifecycle/

[26] Working Group on Information Systems and Services. 2012. *Data life cycle models and concepts: CEOS Version 1.2.* Technical Report. Committee on Earth Observation Satellites (CEOS). http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/White_Papers/WGISS_DSIG_Data-Lifecycle-Models-And-Concepts-v13-1_Apr2012.docx

[27] Alberto Pepe, Matthew Mayernik, Christine L. Borgman, and Herbert Van de Sompel. 2010. From artifacts to aggregations: Modeling scientific life cycles on the semantic web. *Journal of the American Society for Information Science and Technology* 61, 3 (2010), 567–582.

[28] Line Pouchard. 2015. Revisiting the data lifecycle with big data curation. *International Journal of Digital Curation* 10, 2 (2015), 176–192.

[29] Janine Rüegg, Corinna Gries, Ben Bond-Lamberty, Gabriel J. Bowen, Benjamin S. Felzer, Nancy E. McIntyre, Patricia A. Soranno, Kristin L. Vanderbilt, and Kathleen C. Weathers. 2014. Completing the data life cycle: Using information management in macrosystems ecology research. *Frontiers in Ecology and the Environment* 12, 1 (2014), 24–30.

[30] Amir Sinaeepourfard, Xavier Masip-Bruin, Jordi Garcia, and Eva Marín-Tordera. 2015. *A survey on data lifecycle models: Discussions toward the 6Vs Challenges (UPC-DAC-RR-2015–18).* Technical Report. https://www.ac.upc.edu/app/research-reports/html/RR/2015/18.pdf

[31] Carly Strasser, Robert Cook, William Michener, and Amber Budden. 2012. *Primer on data management: What you always wanted to know.* Technical Report. DataONE. https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf

[32] Marianne Swanson, Pauline Bowen, Amy Phillips, Dean Gallup, and David Lynes. 2010. *Contingency planning guide for federal information systems, SP 800-34 Rev.1.* Technical Report. National Institute of Standards and Technology (NIST). https://csrc.nist.gov/publications/detail/sp/800-34/rev-1/final

[33] Barbara M. Wildemuth. 2009. *Applications of Social Research Methods to Questions in Information and Library Science.* Libraries Unlimited.