CHAPTER

THE CENTRALITY OF DATA: DATA LIFECYCLE AND DATA PIPELINES

4

Beth Plale and Inna Kouper

Indiana University, Bloomington, IN, United States

4.1 INTRODUCTION

Any long-distance travel over the roads of the United States will invariably cause a driver to deal with weather of some kind that can disrupt plans. Inclement weather can affect visibility and road conditions, impairing the ability of a driver to drive safely. Approximately 22% of car crashes, nearly 1.3 million annually, are weather-related [1].

Increasing incidents of severe weather attributed to climate change can raise road maintenance costs and increase emergency response times as well. Each year, state and local agencies spend up to 2.5 billion dollars (USD) on road weather maintenance, such as snow and ice removal [2,3]. Choice of treatment and equipment used to clear the roads also depends on weather, with underprediction creating a danger of roads being unsafe and overprediction having a negative impact on the environment.

Uncertainties on the road, transportation disruptions, and economic cost associated with inclement weather present significant challenges for research and practice. The newest sensor and communication technologies can help with some of the challenges as they can deliver real-time traffic and weather data and further increase connectivity, automation, and information density needed for Intelligent Transportation Systems (ITS). According to a recent report from the Intelligent Transportation Society of America (ITS America), 77% of transit companies now have vehicle location and real time arrival data and an increasing number of freeways and arterial roads in the United States are covered by real-time monitoring systems and computerized traffic management systems [4]. Road weather data collection and prediction systems also are getting more and more sophisticated.

As ITS-enabled technologies mature, we can envision many scenarios where intelligent agents provide adaptable, dynamic information needed to make decisions in real time. Such decisions frequently draw on real-time data, such as weather conditions, road temperatures, traffic pluralize, or driver postings of accident locations. But making decisions based solely on data that was created in the last 24 hours can impede decision-making because recent data lacks the historical context needed to see and verify trends. Drawing a line at a certain age of data fails to acknowledge the value of collecting data on slow-changing phenomena. Hike the Rocky Mountains and chances are

Data Analytics for Intelligent Transportation Systems. DOI: http://dx.doi.org/10.1016/B978-0-12-809715-1.00004-3

you are using a topographic map created in the mid-1940s by the United States Geological Survey [5]. Funded with public money through taxes, the datasets from USGS are freely available without fees and have been widely used in many commercial and academic mapping initiatives.

We raise the issue of older data because data is valuable regardless of its age. Old data can in fact be more valuable than new data because it has survived the test of both time and use. Knowing what data to use and how to compare past and present measurements is essential in assessing the impact of phenomena such as climate change or population migration. At the same time, appreciating historical and recent data as similarly valuable requires an understanding of the lifecycle of data. Data is born and dies; it exists on certain media, actively participates in research, or simply sits passively waiting to be used. Taking a "data-first" view on ITS, allows us to think about data independent of its uses and rush evaluations of its relevance. It is this data-first view that allows us to respect data, steward it, and care for it for future generations.

The data lifecycle is an abstract view of what could be called "the life of data objects from birth to death." By data objects here we mean a collection of files and links or a database. The data lifecycle defines a set of stages that not every object goes through. As will be shown in this chapter, there are different views on what are the stages of a data lifecycle. In this chapter we strive to give the reader an understanding of the lifecycle of data. We also make the connection from the data lifecycle to the "data pipeline," the latter being a physical manifestation of portions or all of a data lifecycle through running software.

4.2 USE CASES AND DATA VARIABILITY

Because of the immediacy of weather and road conditions, a reader might be lulled into thinking that only most recent real-time data is useful. In many respects this is right. Data that are gathered from road sensors or any other kind of sensor need to be recent for the users to react in a timely way to an emerging hazard. Use Case I below illustrates how immediate data is used for forecasting and navigation [6]:

> *Case I: Immediate data: Karen gets into her car in Denver, CO to drive north along the Rocky Mountains into Wyoming. While driving she receives a notification on her in-vehicle device "pathcast" tool—a regionally specific map-based prediction that is generated by a new startup company using data from the National Weather Service and the local data from lightning sensors, radars, surface-observing stations, wind profilers, and stream gauges. Karen's pathcast forecast indicates that earlier rains have caused one of the canyon creeks to surge, so Karen turns to her car's sophisticated navigation system to circumvent the flooded roadways.*

For Karen in Case I to be able to make a timely decision, the company or agency that is behind "pathcast" must continuously receive information on degrading weather conditions from the national and local sources mentioned. The company counts on the information not only being recently generated, but also being delivered to the computer systems at the company's headquarters (or their cloud-hosted services) in a timely manner.

There are many cases in transportation informatics and climatology though where data from the past is equally as valuable. Use Case II below illustrates how data from the past 24 hours is used to predict trends.

Case II: Data less than a week old: Storms can linger in an area for many hours causing slowly degrading conditions. A student intern at a weather company is asked to make a short-range prediction of snowfall and its anticipated accumulation. The intern chooses to plot the snow–water ratio on the Y-axis and a mean air temperature measure called "thickness" on the X-axis for a large number of roads in his region over a period of the last 24 hours. He uses least squares regression analysis to create a best-fit line between the points and an equation upon which to predict future degradation. The analysis provides a very localized pattern of weather conditions for further modeling and enriches the intern's experience with weather data.

For the intern in Case II to provide meaningful results in his analysis, data had to be less than a week old. Observed data in this case had to be gathered from the local and state data sources, the most relevant data being the data generated during the period of accumulating snow only—the last 24 hours. As conditions can change dramatically, invalidating the numerical prediction of the plot, the prediction can only be good for some short projection in time beyond the observed measurements. But assuming the best-fit line is a good fit, the intern's approach could reasonably predict short-term future snow conditions.

Finally, there is what might colloquially be called "ancient" or "legacy" data; that is, data older than a week. In an era of seemingly abundant digital data and social media, data older than a few days could be considered simply too ancient to be relevant, but that view can significantly constrain data analysis and the decision maker's lens. As 20th-century philosopher George Santayana famously said, "those who cannot remember the past are condemned to repeat it."[1] Similarly, those who cannot remember or find data from the past, are destined to repeat the previous mistakes and overlook long-term variations.

Case III: "Ancient" data: A researcher wishes to study a portion of a road where asphalt crumbles and cracks more quickly than other parts of the very same road and determine the causes of that. She might want to plot the road topology against known geological data of the region around which the road is built. In this case, the data that she will be using might have been collected 50 or 100 years ago!

The three use cases illustrate the need for understanding and valuing data of varying age. Another important aspect of data that affects its lifecycle is its sheer variability of data types. Decision systems that are built to guide emergency traffic management requires that data arrive in real time from numerous sources as well as on data that is changing less frequently and being collected over the decades. Systems like this may combine any of a number of different data types. The list below describes a few common data sources:

[1] https://www.gutenberg.org/files/15000/15000-h/vol1.html

- Weather and climate data from national and local networks. The US National Oceanic and Atmospheric Administration (NOAA), for example, provides land-based, marine, model, radar, weather balloon, satellite, and paleoclimatic data.[2]
- Sensor data on the road and near the road that collect temperature, humidity, wind speed, air quality, and other measurements. States, for example, deploy Environmental Sensor Stations (ESS) that contain dozens of sensors for various types of measurements [7]. In addition to regular-size sensors, researchers also look into producing very small sensors, the so called "smart dust," which can be deployed in harsh environments [8,9].
- In-vehicle sensor data. Vehicles driving on the road are equipped with a magnitude of sensors that report vehicle speed and position, regular and nonregular conditions, for example, tires slipping or strong braking, and the environment status. Sensor data are streaming data, that is, they are produced continuously, at certain time intervals. The National Center for Atmospheric Research (NCAR) and the US Department of Transportation are working on the development of the Vehicle Data Translator—a tool that can ingest data from many sensors, check it, sort by time, road segment, and grid cell and make the data available for other applications [10].
- Satellite and other observations of land, streams, and other environmental features that affect the roadway. The Geostationary Operational Environmental Satellite system (GOES), operated by the U.S. National Environmental Satellite, Data, and Information Service (NESDIS), for example, is a source of air, cloud, precipitation, and many other kinds of data.[3]
- Social media data. Social media data, such as Twitter data, can be queried and integrated into ITS for language processing and categorization or for network analysis. Such processing can provide additional information about road conditions and driver reactions to the weather as well as predict weather and its impact on traffic [11].

The importance of data to advancements in ITS has been recognized by the U.S. government. The U.S. Department of Transportation launched the Connected Data Systems Program—a vision to create systems that will "operationalize scalable data management and delivery methods exploiting the potential of high-volume multisource data to enhance current operational practices and transform future surface transportation systems management" [12]. A core element of this program is Research Data Exchange (RDE)—a platform that provides access to data from connected vehicles, including transit, maintenance, and probe vehicles, and various types of sensors, such as incident detection systems, traffic signals, and weather and other types of ITS sensors. The goal of RDE is to enable a wide range of ITS-related analysis and research.

To make sure that a new data source (a new instrument, for instance) fits within the increasingly complex data ecology around transportation research and practice, a data manager or researcher must be cognizant of the lifecycle of data. Not only must a data manager be knowledgeable, the data itself must carry its history and lineage with it, like carrying its own passport. That is, for a data source to be suitable for integration into complex modeling and transportation systems decision-making, the data entering an ecosystem must carry with it information about its origin, context, and history of transformations to guide future uses. Data that cannot be accessed and verified has little chance to have a long life and gets quickly forgotten.

[2] https://www.ncdc.noaa.gov/data-access

[3] http://www.goes-r.gov/products/overview.html

Today data are being assembled from a growing number of sources to provide a deeper and broader picture of emerging conditions in different domains. In order to reach back in time to harvest even more resources for explanations of what is happening and for predictions of what will happen, one must understand the data lifecycle. In the next section, we look at various models of the data lifecycle and how they address these and other data management issues.

4.3 DATA AND ITS LIFECYCLE

It is helpful to think about data management within a framework that records actions applied to or with data as the actions are taken rather than after the fact. Models that describe data objects (e.g., collections of files and links or a database) through a set of time-ordered stages are called lifecycle models. Many lifecycle models exist, including domain-specific, regional and industry-specific models [13,14]. Below we describe models that are generic data lifecycle models, that is, they can be applied to different domains and adapted as needed.

4.3.1 THE USGS LIFECYCLE MODEL

The U.S. Geological Survey (USGS) developed the science data lifecycle model "to facilitate shared recognition and understanding of the necessary steps to document, protect, and make available the Bureau's valued data assets" [15]. The model consists of primary and cross-cutting elements, with the former representing a linear sequence of steps that can be used iteratively and the latter representing activities that need to be used continuously (see Fig. 4.1).

The primary elements include the following steps: plan, acquire, process, analyze, preserve, and publish/share. Each element is briefly described below.

Planning is a stage during which project participants consider all activities related to the project's data and decide how they are going to approach them. During this stage, the project team considers implementation of subsequent stages as well as the resources they will need and the

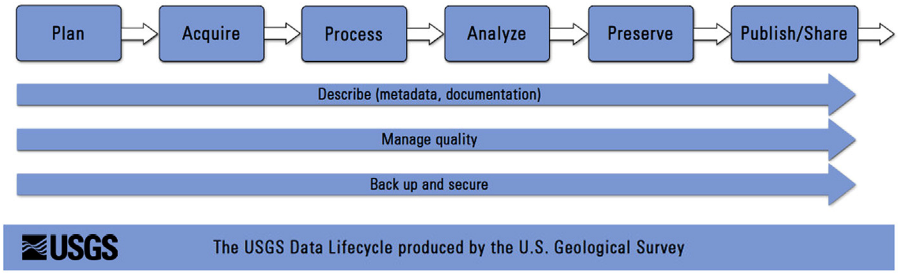
Plan
Acquire
Process
Analyze
Preserve
Publish/Share
Describe (metadata, documentation)
Manage quality
Back up and secure
USGS
The USGS Data Lifecycle produced by the U.S. Geological Survey

FIGURE 4.1

USGS data lifecycle model.

intended outputs for each stage of the data lifecycle. The model recommends to create a data-management plan as an outcome of the planning stage.[4]

Acquiring includes activities through which data is retrieved, collected, or generated. National Weather Service data or satellite observations are examples of data that can be acquired from external sources. Sensor data can be collected by installing and monitoring sensors. At this stage it is important to consider relevant data policies and best practices that ensure data integrity. The outcome of this stage is the project design and a list of data inputs.

Processing represents activities associated with further preparation of collected data for analysis. It may entail design of a database; integration of disparate datasets; loading, extraction, and transformation of data. During the processing stage it is important to consider standards and tools that are available for data storage, processing, and integration. The outcome of this stage is datasets that are ready for integration and analysis.

The *analysis stage* is when the exploration and interpretation of data and hypotheses testing takes place. Analytical activities can include summarization, statistical analysis, spatial analysis, modeling, and visualizations and are used to produce results and recommendations. Proper data management during this stage improves the accuracy and efficiency of data analysis and provides a foundation for future research and its application. The outcome of this stage is a formal publication or a report.

The next two stages—*preservation* and *publication* of data have been recently recognized as elements of data lifecycle that are as important as data collection and analysis. Preservation refers to preparing data for long-term storage and accessibility. Publication may include references to data in peer-reviewed publications or dissemination of data through web sites, data catalogs, and specialized repositories. The USGS model intentionally puts preservation ahead of publication to emphasize that federally funded research must plan for the long-term preservation of data, metadata, and any additional products and documentation and thus such planning is as important as publishing research results. The goal of having data publishing as part the data lifecycle is to remind that data is a research product that is equal to other outputs of research, such as papers or presentations.

Critical cross-cutting activities in the USGS data lifecycle model include *describing*, that is, providing metadata and data documentation, *managing quality*, that is, implementing quality assurance measures and undertaking ongoing quality control, and *backup and security*, that is, taking steps to manage risks of data corruption and loss. These activities need to be performed throughout the project timeline, and they facilitate better quality, understanding, and future uses of the data.

4.3.2 DIGITAL CURATION CENTER (DCC) CURATION MODEL

The data curation lifecycle model developed at the UK Digital Curation Center (DCC) focuses on steps necessary for successful curation and preservation of data. Its goal is to provide a generic framework for managing digital objects so that it can be adapted to different domains and types of materials [16]. The model represents a cycle of three types of actions—the full lifecycle actions, the sequential actions, and the occasional actions that occur as needed (see Fig. 4.2).

[4] See https://www.dataone.org/sites/all/documents/DMP_MaunaLoa_Formatted.pdf for an example of a data-management plan.

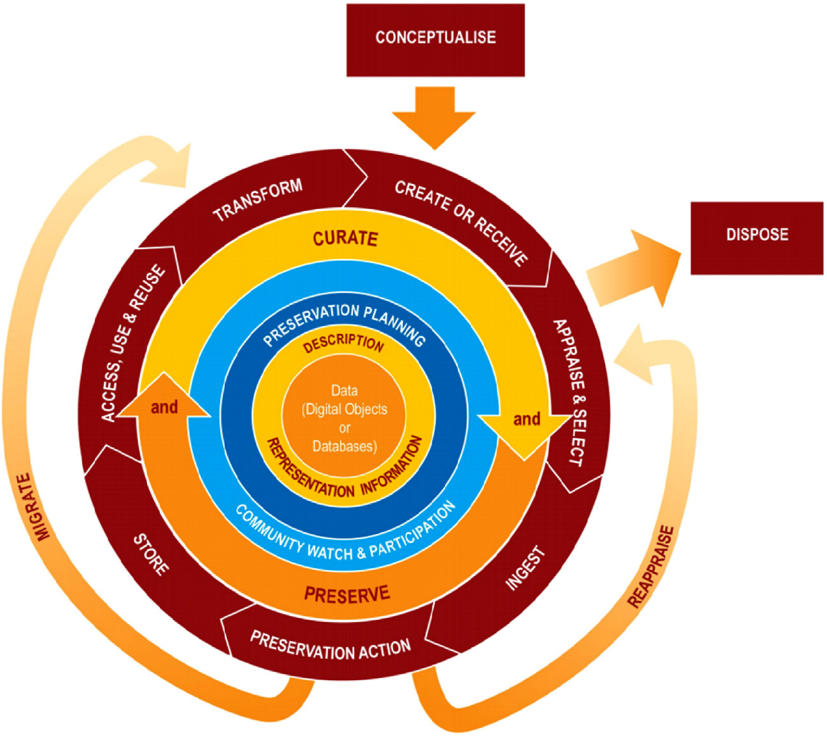

FIGURE 4.2

The DCC digital asset lifecycle model.

Full lifecycle actions are activities that need to be performed throughout the lifecycle of digital objects. There are four full lifecycle actions in the DCC model (represented as three closes circles around the data circle)—description and representation, preservation planning, community watch and participation, and curation and preservation. *Description and representation* refers to actions of providing metadata that is necessary to describe a digital object so that it can be accessible and understandable in the long-term. *Preservation planning* includes plans for administration of all the actions necessary for curation. *Community watch and participation* emphasizes the importance of relying on the existing community tools and standards as well as the importance of contributing to their development. *Curate and preserve* actions raise awareness about the need to promote curation and preservation throughout the lifecycle of digital objects.

The sequential actions follow the trajectory of a project—from *conceptualization* and data collection to making decisions about what needs curation and preservation (*appraise and select*) to steps that are needed to make digital objects available for long-term access and reuse (*ingest, preservation action, store, access, use & reuse*). The final step "*transform*" refers to the reuse stages of digital objects when new and derived data products can be created from the original. At every stage the DCC lifecycle model emphasizes creation and registration of appropriate metadata, ensuring data authenticity and integrity, and adhering to relevant community standards and tools. Occasional actions include *disposition* of data, that is, transfer to archives and destruction, *reappraisal*, that is, verification in case validation procedures have failed, and *migration*, that is, transformation of data to a different format to make it immune from hardware or software obsolescence.

While this model does not place stages of research at its center, it recognizes that knowledge production and enhancement are essential components of the curation lifecycle model [17]. As scientific research generates new knowledge about the world and encodes it in digital resources, the products of such research need to be organized and codified for efficient consumption by humans and machines. The model implies that metadata should include not only text annotations, but also rules and formalized entities that can be used by automated services and tools. The efforts of managers and curators of data contribute to knowledge enhancement and should be part of digital object metadata as well.

4.3.3 DATAONE MODEL

The DataONE data lifecycle model is developed within a National Science Foundation (NSF) funded project on the development of tools and services to support data management and sharing in the environmental sciences [18]. It provides an overview of the stages involved in successful management and preservation of data for use and reuse. The model includes eight generic components that are part of any project that includes the use of data, but it can be adapted to various domains or communities (see Fig. 4.3).

The DataONE model is similar to the USGS lifecycle model as it includes similarly titled components—plan, collect, assure, describe, preserve, discover, integrate, analyze—but the interpretations are slightly different.

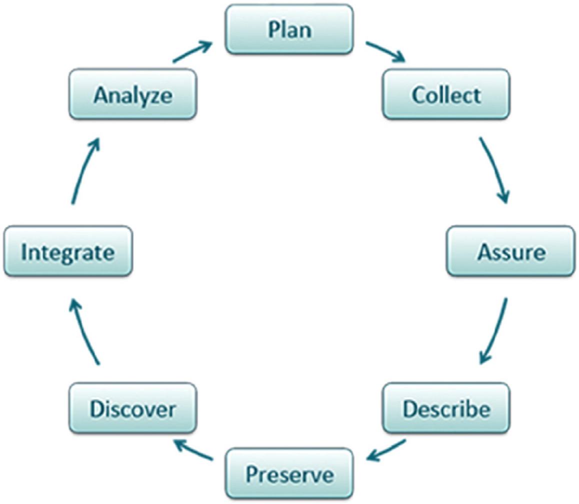

FIGURE 4.3

DataONE data lifecycle model.

Planning includes decisions with regard to how the data will be collected, managed, described, and made accessible throughout its lifetime. Collecting refers to activities of gathering data using observations, instrumentation, and other methods. Assuring refers to guaranteeing the quality of the data through checks and inspections that are accepted within a particular research community. Describing includes recording all relevant technical, contextual, administrative, and scientific information about data (metadata) using the appropriate metadata standards. Preserving includes preparing and submitting data to an appropriate long-term repository or data center. Discovering refers to activities that are aimed at providing tools and metadata approaches for finding and retrieving data and using such tools for locating and using data in one's research. Integrating refers to accessing multiple data sources and combining data to form datasets that can be analyzed within the specific set of research questions or hypotheses. Finally, analyzing involves using various tools, methods, and techniques to analyze data.

The DataONE model encompasses the whole cycle, but it does not require all activities to be represented in every project. Some projects might use only parts of the lifecycle, although quality assurance, description, and preservation activities are crucial to any project.

4.3.4 SEAD RESEARCH OBJECT LIFECYCLE MODEL

The SEAD Research Object lifecycle model was developed by the authors of this chapter as part of the National Science Foundation-funded project that aimed at developing tools to support data curation, publishing, and preservation [19]. This model draws on the concept of Research Object [20–22] and work on provenance [23–24] and provides a framework for easier dissemination and reuse of digital units of knowledge. According to the Research Object (RO) approach, sharing and reproducibility can be best supported if research resources are aggregated into bundles that are structured and contain information about resources' lifecycle, ownership, versioning, and attribution.

RO is a bundle that encapsulates digital knowledge and serves as a vehicle for sharing and discovering reusable products of research. The principles of ROs that have been then implemented in the SEAD model include: aggregation (content), persistent identification, attribution (links to people and their roles), and provenance (references to states and relationships). Therefore, RO is defined through five interrelated components (see Fig. 4.4):

- a *unique ID*, a persistent identifier that is never reassigned;
- *agents*, information about the people who have touched the object in important ways (e.g., creator, curator, and so on);
- *states*, which describe where in its lifecycle an RO currently is;
- *relationships*, which capture links between entities within the object, such as datasets, processing methods, or images, and to other ROs, and
- *content*—the data and related documents.

To describe the lifecycle of ROs and support their integration and reusability, this model focuses on the behavior of a research object as it passes through the creation–analysis–publish–reuse cycle. The model still fits within the regular research lifecycle, but it also allows to capture relationships between research objects as they get created or acquired, prepared for publication, derived from one

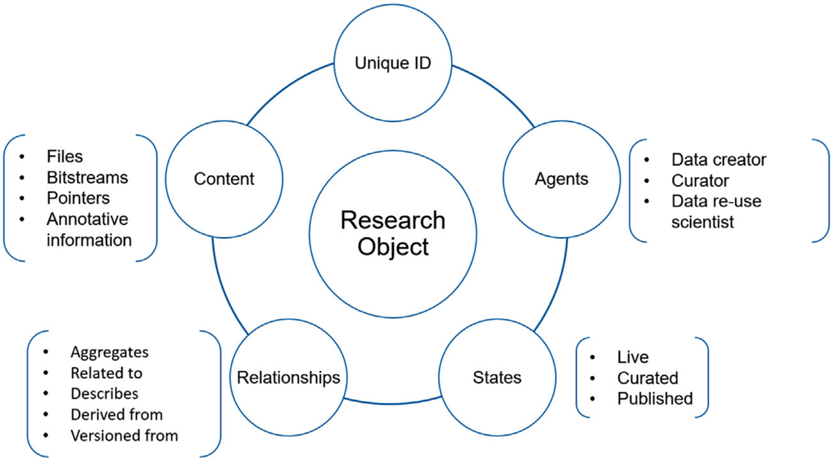

FIGURE 4.4

SEAD research object as a main concept in the lifecycle model.

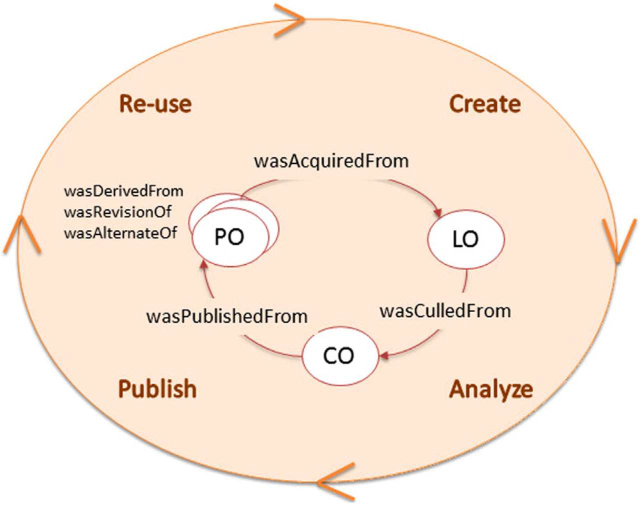

FIGURE 4.5

SEAD RO model.

another, replicated, and so forth. Such a model serves as the basis upon which software can be written to track RO changes and movements in a controlled and predictable manner.

The model has two parts: (1) *states*, which define the condition of a data bundle as it goes through research stages, and (2) *relationships*, which capture the relationship *between ROs*. Fig. 4.5

below provides an overview of how an RO transitions through research, and how it relates to its derivatives once published.

As can be seen from Fig. 4.5, in addition to providing a high-level overview, this model offers a formalized and technical specification of how ROs move through research and how they can be tracked by systems and services. The states and relationships are drawn from two enumerated sets, where states represent conditions of going through the stages of data collection and analysis (Live), selection and description (Curation), and dissemination and reuse (Publication). The relationships are a subset of the properties that can be formally defined by a provenance ontology such as PROV-O or a publishing ontology[5]:

- *States = {Live Object (LO), Curation Object (CO), Publishable Object (PO)}*
- *Relationships = {wasAcquiredFrom, wasCulledFrom, wasDerivedFrom, wasSharedWith, . . ., wasRevisionOf, alternateOf}*

An RO exists in one of the three states: as a Live Object (LO), as a Curation Object (CO), or as a Publishable Object (PO). As data management is rarely the focus of the beginning stages of research, we consider the active data as existing in the "Wild West"—a space where organization and descriptions are loose, even though it is crucial to continue raising awareness and encouraging good practices of data management among the data contributors. Such space can exist on multiple computers, within tens or thousands of folders and with inconsistent and confusing names.

A researcher or a practitioner acquires data from various sources, through his/her own effort and by retrieving the existing data. This data is the LO and it can be related to its sources via the "wasAcquiredFrom" relationship. Then the researcher identifies subsets necessary for processing and analysis and culls those subsets from the larger datasets. Over time, the researcher will prune and organize the material to be suitable for analysis, interpretation, formal representation, and subsequent publication. This culled content becomes the CO, an object related to its LO by the "wasCulledFrom" relationship.

We consider the CO as existing in a more controlled setting, in the "Boundary Waters," where changes are limited and data and metadata are verified and properly recorded. Once the researchers and data curators agree that the content and descriptions of the research product are ready, the RO can move to a new state as a PO.

The PO is a data product that contains everything so that it can be deposited into a repository or shared with other researchers. It can be related to the CO by the "wasPublishedFrom" relation. The PO exists in the "Control Zone" where all actions have to be tracked. Particularly, the actions of correction, sharing, derivation, duplication, and other forms of use and reuse. These actions form the past and future lineage of a family research objects and facilitate integration of digital objects without loss of quality or context. As the lineage gets recorded, over time the model will result in a network of links between research objects, creating a genealogy network for published scientific data.

All four data lifecycle models that are introduced in this chapter provide a high-level overview of the data lifecycle with a varying amount of detail about what could or should be done at every stage with regard to data management. Table 4.1 below provides a comparison between models and their stages.

[5] http://www.sparontologies.net/

Table 4.1 Comparison of Data Lifecycle Models

	USGS	DCC	DataONE	SEAD
Sequential stages	Plan	Conceptualize	Plan	
	Acquire	Create	Collect	Create
		Appraise & select	Assure	
	Process	Ingest	Describe	
	Analyze			Analyze
	Preserve	Preservation action	Preserve	
	Publish/Share	Store	Discover	Publish
		Access, use, & reuse	Integrate	
		Transform	Analyze	Reuse
Cross-cutting or complementary aspects	Metadata	Metadata	Metadata	Metadata
	Quality	Preservation	Quality	Provenance
	Security	Community	Preservation	Curation
		Curation		

Some models, such as the USGS and the DataONE models, are useful in steering data creators and users toward good data management and encouraging them to search for appropriate metadata schemas and tools for sharing data. The DCC model may be particularly useful for data curators as it provides more details about what to do after the data has been collected and analyzed. The SEAD model provides a simple mechanism for tracking data movements and for ensuring integrity of data.

Data lifecycle models are frequently built into software services and policies that are used in practice. These software services take the form of data pipelines. In the next section, we describe the more hands-on aspects of the data lifecycle in practice.

4.4 DATA PIPELINES

The sequence of increasingly historical data presented as Use Cases in Section 4.1 of this chapter illustrates that while it is the most recent data that often grabs our attention because it captures what is happening now, there is considerable value in data that are older. Hence when we talk about data lifecycles, we need to be cognizant of the fact that data used to solve problems in transportation informatics may be 5-minutes old, 5-days old, or 50-years old.

In today's connected world, real time data and historical data frequently reside in any number of databases or data repositories. Computationally intensive ITS research became mainstream in transportation engineering research and innovations due to the recent advancements in instruments, in-vehicle and roadside sensor technologies and networks, wireless communication between vehicles and transportation infrastructure, Bluetooth, video cameras, image recognition, and a whole host of other technology that has gotten lighter, cheaper, and more reliable. In addition, future automated vehicles will generate massive amounts of sensor data which are thousands times larger in volume than today's most advanced vehicle models. The data from this plethora of real-time and less than real-time sources has to be directed to locations where it can be processed.

Moving data around so that it is where it needs to be and when is handled by software systems, or connected software services that communicate with one another. This data routing depends on the presence of fast Internet connectivity to the multiple destinations, possibly world-wide, that receive the data. Suppose the transportation data for a major city in the United States is distributed by prior agreement simultaneously and in real time to a handful of institutions, including universities and government labs. In the discussion below a hypothetical university is considered which receives this data in real time. And in doing so a type of software framework is discussed which can guide the researchers on how to turn the data from the plethora of sensors and instruments that characterize transportation informatics to scientific or practical insight.

The "data pipeline" is an abstract way of talking about the data handling components written in software that are applied to data objects in sequence. The data pipeline is a useful abstraction because it helps one to think about (a) how data are pushed in real time from sensors and instruments through processing steps towards outcomes and (b) how to optimize data handling while minimizing its cost. A data pipeline thus is an abstraction for managing and streamlining data processes throughout the data lifecycle.

"Workflow" is another concept similar and related to "data pipeline" that describes automated data handling [25]. A workflow is a set of software tasks that are orchestrated or automatically run by a workflow engine. Frequently the workflow is defined using an abstract planning language. The plan is then fed into a workflow engine which will then run the workflow according to the plan. The workflow and the data pipeline share much in common from a software architecture point of view. A data pipeline could be built from a workflow system but as regular and repeatable as the data pipeline is, the workflow orchestration engine is seen as too cumbersome for implementing a data pipeline.

The tasks in a data pipeline could be manual tasks, carried out by humans, or they could be automated and triggered periodically by scripts that run all the time. The tasks may be individual or interchangeable enough so that that the order of their execution does not matter. Alternatively, the order may be extremely important because the tasks depend on where in the pipeline they run. Commonly, there is an accepted order to the execution.

In practice, data pipelines often facilitate the beginning of the data lifecycle, to define how the data, once created, is processed for use. For instance, NASA has defined a policy for how the data from its satellites and other instruments are processed through their data-processing level document [26]:

> Data products are processed at various levels ranging from Level 0 to Level 4. Level 0 products are raw data at full instrument resolution. At higher levels, the data are converted into more useful parameters and formats. All [. . .] instruments must have Level 1 products. Most have products at Levels 2 and 3, and many have products at Level 4.

Data pipelines can also occur later in the data lifecycle, for instance, at the point where a data object moves from storage to compute or when the object is ready for more broadly sharing outside the environment in which it was first collected. Two practical examples below, one at the beginning of the data lifecycle and another at the publishing point, help to better understand how a data pipeline works.

A transportation manager for Denver, Colorado, a major metropolitan area, wants a decision support system to respond to road weather conditions. Denver is known for its highly variable wind patterns and weather caused by the Rocky Mountains just 20 miles West of downtown Denver. The transportation manager charges the IT director with setting up a data pipeline to feed data in real time to the decision support system located in the traffic management center (TMC) in Denver and used by the transportation manager.

The IT director begins by sketching out the data pipeline shown in Fig. 4.6.

The data pipeline in this example is considerably simplified to bring out data-related issues. The figure shows on the left four different data-generation sources. The first and third are environmental sensor networks. Each network is a set of five road temperature sensors. The IT director intends to deploy one sensor network on interstate I-70 just east of Dillon, Colorado right where traffic enters and exits the Eisenhower Tunnel. This portion of I-70, which drops under the Continental Divide, is famous for being highly dangerous for drivers under bad weather conditions. The second unit will be deployed in I-70 near Genesee, Colorado, just into the mountains on the boundary of the Denver metropolitan area.

The temperature sensors will communicate with a central receiving unit, a small, embedded computer that is co-located with the sensors. The sensors will take readings once every 5 seconds and send their data to the embedded computer. The embedded computer will receive five readings every 5 seconds (60 readings a minute) and send them to the Denver TMC. Depending on the software and hardware power, the embedded system can do more than simply transmit data. It can be tasked with data preprocessing and synthesis to create dynamic geographic temperature maps, with each map including a reading from each sensor, and sending the maps to the TMC at a rate established by the manager's needs, for example, daily summary maps or frequent update maps of 12 per minute. The TMC will receive the data from the embedded computer and store it for further processing, including validating and storing raw and synthesized data, checking the maps for quality, and adjusting the timestamp from Mountain Time to UTC so that time representation is uniform and can be integrated with sensor data from other regions.

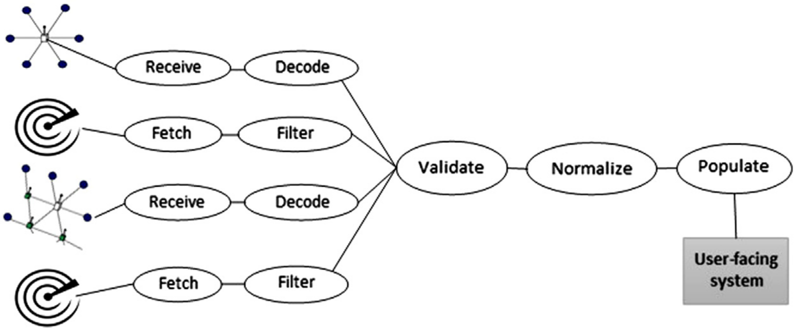

FIGURE 4.6

Weather data pipeline.

In addition to environmental sensors, the IT director deploys two Collaborative Adaptive Sensing of the Atmosphere (CASA) radars, in each of the same two locations, to improve forecasting speed and accuracy [27]. CASA[6] radars, called Distributive Collaborative Adaptive Sensing (DCAS), are low-cost, small-footprint Doppler radars that operate at short-range with good resolution and coverage throughout the lower atmosphere and produce readings in 1-minute intervals [28]. In the example above, the CASA data, once it arrives at the TMC in Denver, is filtered to discard all products except one, which the transportation manager deems important to predicting his department's response to an upcoming storm.

The data pipeline in the CASA example can be implemented as one large data pipeline or as five data pipelines. In the latter case, there will be one small pipeline per instrument, then another pipeline that fuses the instrument pipelines into a single temporally consistent pipeline. By being temporally consistent, the events are ordered and coordinated in time—a serious challenge in data-processing, considering that there are numerous clocks involved, and that using batteries in the instruments could mean that the clocks of the instruments can slow down. The speed of the Internet connection is another challenge as it can create slack in the arrival rate of the data at the central pipeline and undermine real-time decision-making.

This hypothetical example illustrates a data pipeline for ingesting observational (sensed) data in real time from multiple sources. In present days, the pipeline will likely be built using Java or Python, a relational database such as MySQL or a noSQL databases such as MongoDB to store the events. The event format would be JSON or even JSON-LD as the latter provides ways to encode semantic information as needed. The service would likely have one or more RESTful APIs—a set of programmatic components that conform to the constraints of the Representational State Transfer (REST) architecture that enables standardized querying and pulling results from the database regularly as needed to populate its user-facing systems.

The second example is a pipeline that is used when data needs to be published. The transportation manager likely does not have an obligation to publish any data from their work in protecting roads and lives. However, sharing data goes way beyond formal academic publications. Managers may want to have a space where they can exchange data with other cities and states or with the public through crowdsourcing initiatives. Some of the data the manager uses may come from government sources that have to be made available. Researchers who may work on improving predictive modeling techniques and use the data from the sensors and radars from our first example need to think about publishing their data, especially if research is federally funded [29].

The second pipeline example was drawn from an existing data pipeline supported as part of the SEAD project and from a collective profile of data communities supported by SEAD.

Two environmental sensor networks and radar systems deployed along I-70 road in Colorado generated a large amount of data that became of interest to a group of atmospheric and data scientists who received a grant to develop a next-generation numerical weather prediction system designed for both atmospheric research and traffic operational forecasting needs. While the researchers have facilities for computation, they lack infrastructure and tools for subsequent selection of data objects for sharing beyond their immediate team.

[6] www.casa.umass.edu

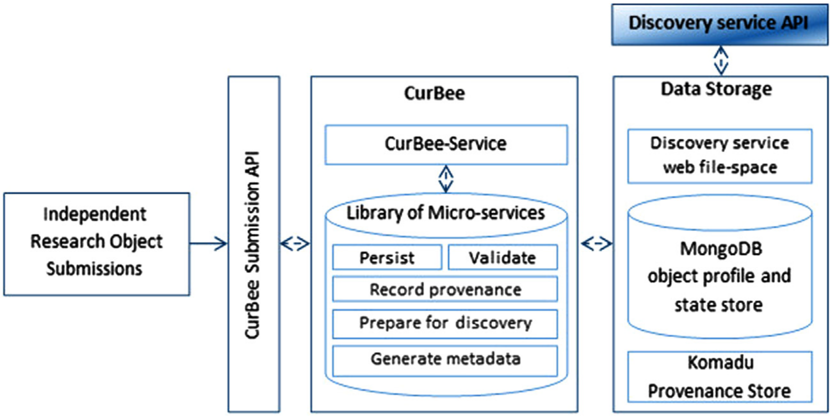

FIGURE 4.7

Curbee pipeline architecture.

The data pipeline for preparation, selection, and subsequent sharing of data that can help in the example above is implemented by a software framework called *Curbee* [30]. Curbee is a part of the larger software implementation that transforms an abstract SEAD data lifecycle model into a working system of software- and human-supported tasks, such as storing and pruning active data, managing access of various team members, creating metadata, and publishing and preserving data.

Curbee is designed to be a loosely coupled component, that is, it can work on its own or be incorporated into a larger pipeline. It accepts data objects combined into a bundle or single package and utilizes a set of microservices to move the object through the pipeline (see Fig. 4.7).

The microservices include the following:

- object validation—making sure that the bundle is complete and its content corresponds to the declared description
- persistence—storing the object and its state in the database
- provenance—registration of the object's lineage through the relationships of "DerivedFrom," "AcquiredFrom," and so on[7]
- preparation for discovery—making data object compliant with data discovery services such as DataONE[8], and
- metadata generation—creating metadata according to the schemas required by publishers and discovery services.

After performing the microservices, Curbee notifies a repository recommended for this object of the prepared submission and synchronizes metadata with data discovery services.

[7] See "SEAD Research Object Lifecycle model" in Section 4.3 above.

[8] https://www.dataone.org/

Curbee uses Java programming language and other present-day technologies and standards, such as the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) and BagIt protocols,[9,10] serialization through JSON-LD, and noSQL database MongoDB. As the object goes through the data publishing pipeline, status messages and timestamps are added to the package to make the pipeline more transparent to both users and software agents.

The first and second data pipeline examples are a small illustration of the myriad of tasks that can be handled by a data pipeline. In addition to data ingest and data publication, pipelines can be designed to handle the following:

- Data cleaning and quality assurance: use manual and automated procedures to detect erroneous, invalid, or inconsistent data and correct it.
- Data assurance/curation: use standardized schemas to add metadata, use ontologies and provenance techniques to add semantic meaning and record and track lineage.
- Data analysis and visualization: use statistical, text/data mining and visualization techniques to make sense of data and phenomena they describe.
- Data discovery: use search, access, and retrieval tools to find datasets.
- Data fuse and integration: integrate multiple data and knowledge into a consistent and useful representation that can provide deeper and bigger insights.
- Data publication: make data available to others via publishing datasets that have a persistent ID and a use license.
- Data preservation: prepare data for long-term storage in an archive and ensure it can be used in the future.

4.5 FUTURE DIRECTIONS

The ITS is a fast-developing new area that is being shaped by rapid technological advances as well as the development of new tools and algorithms. Cheaper sensors, distributed computing, and artificial intelligence techniques will inevitably reshape how we travel and navigate. The centrality of data in these changes is obvious, and we need to think deeper about these changes to be able to adapt the data lifecycles and pipelines to the changing needs and emerging capabilities of ITS. The following research directions can be identified as important in the nearest future:

- Data lifecycle models for better decisions. Most models currently identify general steps in collecting and managing data, however, as the systems and analyses are getting more complex, how can we further specify and adapt the lifecycle models to make better decisions? How can we balance machine and human resources and existing possibilities and constrains? How can our models accommodate shifts in decision-making toward more automation and artificial intelligence?
- Data pipelines for global distributed computing and Internet of Things. As Internet of Things, that is, data from many devices and objects around, becomes more and more ubiquitous, it

[9] https://www.openarchives.org/ore/
[10] https://wiki.ucop.edu/display/Curation/BagIt

poses a challenge to capturing how the data is created and how it moves through the stages of data lifecycle. How can many various devices be tracked uniformly? Will the data be captured in real time or in batches? Where and how the data will be stored and analyzed?

- Academic, government, and commercial data integration and management. This is particularly important in the areas of automated vehicles, where much of the research and innovation is currently led by commercial companies. How can the data lifecycle and pipeline models accommodate integration of proprietary data and provide mechanisms for flexible data sharing and exchange?
- Data for improved access and efficiency. Our data techniques need to consider the existing and potential inequalities in access to data, particularly from the disadvantaged groups. How can we preserve and share data and design algorithms for a better, more just world that provides access to intelligent transportation for all and minimizes risks of exclusion as well as environmental pollution? And what ethical protections and legislative control are needed to enable it?

4.6 CHAPTER SUMMARY AND CONCLUSIONS

This chapter discusses data lifecycle and data pipeline as two related concepts that together provide researchers and practitioners with a useful framework that places data at the center of science and decision-making and offers a holistic approach to managing data in both planning and implementation. Through illustration using use cases and examples we give tangibility to the role that data plays in building ITS and in advancing science and society.

Where data lifecycles are a framework for thinking about the stages through which a data object passes in its life, the data pipeline is a practical notion often constructed as a set of tools, services, and APIs that support the data lifecycle and help to optimize data processes. The data lifecycle approach can be useful for those working with weather and transportation data in thinking about how they are going to gather their data, improve its quality and consistency, and make sure data can be connected with other data, past, present, or future. Data pipelines are employed where processes are repeatable and can be automated, which is an essential and fast-developing part of connected transportation systems.

4.7 EXERCISE PROBLEMS AND QUESTIONS

4.7.1 EXERCISE 1. DEFINING AND DESCRIBING RESEARCH DATA

Discuss your research project and research data in small groups. Think about the following questions:

- What is your research topic and research “location”?
- What physical data will you work with, for example, soil samples, water measurements, etc.?
- What types of data will you obtain or create digitally, for example, from social media?
- What is the origin of your data, for example, data from government or commercial instruments, location samples, published sources, etc.?
- Where will your data end up after the project?

- How will you look after your data?
- Any there any other issues for management and curation of your digital data? For example, risks, ownership, ethical issues?

4.7.2 EXERCISE 2. MAPPING RESEARCH PROJECT ONTO THE LIFECYCLE

Using one of the lifecycle models discussed in this chapter, describe your most recent project that involved working with data. Think about the following questions:

- What was the most important cross-cutting activity in your project? For example, generate high-quality data, share the data with others, help others use this dataset, etc.
- What stages of the data lifecycle took most of your time and effort?
- What stages were missing from your project?
- What storage architectures may be appropriate throughout the lifecycle? How will your architecture change as you go from stage to stage?
- How could the lifecycle model help you improve your project?

4.7.3 EXERCISE 3. DATA ORGANIZATION

A systematically organized data is very important for future analysis. When you work with data every day, its organization is obvious to you, but it may be hard to understand to others who know nothing about the project. A good logical system of data organization helps to share and exchange data. When deciding how to organize your data, think about the nature of data. In ITS-related projects, data maybe organized by instrument or location of where data is coming from. It is also possible to organize chronologically, especially when you work with both real-time and historical data.

Select a type of data you are most familiar with or would like to work with in the future (see examples of ITS data provided in this chapter). Think about how you would organize your project if this type of data was your primary data. Consider the following:

- What is the most appropriate logic for this type of data, for example, instrument, location, time, or something else?
- How would you organize and store data documentation and software needed to process the data?
- What other materials might be important to organize and store along with your data?

4.7.4 EXERCISE 4. DATA PIPELINES

Read about these examples of robust data pipelines http://highscalability.com/blog/2014/3/24/big-small-hot-or-cold-examples-of-robust-data-pipelines-from.html. In small groups discuss what makes these examples robust and what are the differences and similarities across these pipelines. Consider the following:

- Structured versus unstructured data stores
- Data normalization
- Data integration and provenance

REFERENCES

[1] USDOT (US Department of Transportation), "How Do Weather Events Impact Roads?," 2016. [Online]. Available: http://www.ops.fhwa.dot.gov/weather/q1_roadimpact.htm.

[2] USDOT (US Department of Transportation), "Disbursements by States for State-administered, classified by function," 2009. [Online]. Available: http://www.fhwa.dot.gov/policyinformation/statistics/2007/sf4c.cfm.

[3] USDOT (US Department of Transportation), "Disbursements For State-administered Highways – 2014 Classified By Function," 2015. [Online]. Available: http://www.fhwa.dot.gov/policyinformation/statistics/2014/sf4c.cfm.

[4] ITS (Intelligent Transportation Society of America), "Annual Report 2010–2011," 2011.

[5] L. Moore, "US Topo — A New National Map Series," Directions, 2011. [Online]. Available: http://www.directionsmag.com/entry/us-topo-a-new-national-map-series/178707. [Accessed: 16-Jul-2016].

[6] Committee on Weather Research for Surface Transportation, Where the Weather Meets the Road: A Research Agenda for Improving Road Weather Services, The National Academies Press, Washington, DC, 2004.

[7] J. Manfredi, T. Walters, G. Wilke, L. Osborne, R. Hart, T. Incrocci, et al., "Road Weather Information System Environmental Sensor Station Siting Guidelines," Washington, DC, 2005.

[8] J.M. Kahn, R.H. Katz, and K.S.J. Pister, "Mobile networking for smart dust," in *Fifth ACM Conf. on Mobile Computing and Networking (MOBICOM)*, 1999.

[9] D. Chawla and D.A. Kumar, "Review Paper on Study of Mote Technology: Smart Dust," in National Conference on Innovations in Micro-electronics, Signal Processing and Communication Technologies (V-IMPACT-2016), 2016.

[10] S. Drobot, M. Chapman, B. Lambi, G. Wiener, and A. Anderson, "The Vehicle Data Translator V3.0 System Description," 2011.

[11] L. Lin, M. Ni, Q. He, J. Gao, A.W. Sadek, Modeling the Impacts of Inclement Weather on Freeway Traffic Speed, J. Transp. Res. Board 2482 (2015).

[12] D. Thompson, "ITS Strategic Plan—Connected Data Systems (CDS)," 2014.

[13] A. Ball, "Review of Data Management Lifecycle Models," 2012.

[14] Y. Demchenko, Z. Zhao, P. Grosso, A. Wibisono, and C. de Laat, "Addressing Big Data challenges for Scientific Data Infrastructure," in 4th IEEE International Conference on Cloud Computing Technology and Science Proceedings, 2012, pp. 614–617.

[15] J.L. Faundeen, T.E. Burley, J.A. Carlino, D.L. Govoni, H.S. Henkel, S.L. Holl, et al., "The United States Geological Survey Science Data Lifecycle Model," 2013.

[16] S. Higgins, The DCC Curation Lifecycle Model, Int. J. Digit. Curation 3 (1) (2008) 135–140.

[17] P. Constantopoulos, C. Dallas, I. Androutsopoulos, S. Angelis, A. Deligiannakis, D. Gavrilis, et al., DCC&U: An Extended Digital Curation Lifecycle Model, Int. J. Digit. Curation 4 (1) (Jun. 2009) 34–45.

[18] W. Michener, D. Vieglais, T. Vision, J. Kunze, P. Cruse, G. Janée, DataONE: Data Observation Network for Earth — Preserving data and enabling innovation in the biological and environmental sciences, D-Lib Mag. 17 (1/2) (Jan. 2011).

[19] B. Plale, R.H. McDonald, K. Chandrasekar, I. Kouper, S.R. Konkiel, M.L. Hedstrom, et al., SEAD virtual archive: Building a federation of institutional repositories for long-term data preservation in sustainability science, 8th International Digital Curation Conference (IDCC-13) 8 (2) (2013) 172–180.

[20] K. Belhajjame, O. Corcho, D. Garijo, J. Zhao, P. Missier, D. Newman, et al., Workflow-Centric Research Objects: First Class Citizens in Scholarly Discourse, in *Workshop on the Semantic Publishing (SePublica 2012)*, 2012.

[21] S. Bechhofer, D. De Roure, M. Gamble, C. Goble, and I. Buchan, Research Objects: Towards Exchange and Reuse of Digital Knowledge, Nat. Preced., [Online]. Available: http://eprints.soton.ac.uk/268555/.

[22] S. Bechhofer, I. Buchan, D. De Roure, P. Missier, J. Ainsworth, J. Bhagat, et al., Why linked data is not enough for scientists, Futur. Gener. Comput. Syst. 29 (2) (2013) 599−611.

[23] Y. Simmhan, B. Plale, D. Gannon, A Framework for collecting provenance in data-centric scientific workflows, IEEE Int'l Conference on Web Services (ICWS'06), IEEE Computer Society Press, 2006, pp. 427−436. Available from: http://dx.doi.org/10.1109/ICWS.2006.5.

[24] B. Plale, I. Kouper, A. Goodwell, I. Suriarachchi, Trust threads: Minimal provenance for data publishing and reuse, in: Cassidy R. Sugimoto, Hamid Ekbia, Michael Mattioli (Eds.), Big Data is Not a Monolith: Policies, Practices and Problems, MIT Press, 2016.

[25] D. Gannon, B. Plale, S. Marru, G. Kandaswamy, Y. Simmhan, S. Shirasuna, Dynamic, adaptive workflows for mesoscale meteorology, in: I.J. Taylor, E. Deelman, D.B. Gannon, M. Shields (Eds.), Workflows for e-Science, Springer, London, 2007, pp. 126−142.

[26] NASA, "Earth Science Data Processing Levels," 2016. [Online]. Available: <http://science.nasa.gov/earth-science/earth-science-data/data-processing-levels-for-eosdis-data-products/>.

[27] B. Plale, D. Gannon, J. Brotzge, K. Droegemeier, J. Kurose, D. McLaughlin, et al., CASA and LEAD: adaptive cyberinfrastructure for real-time multiscale weather forecasting, Computer (Long. Beach. Calif). 39 (11) (Nov. 2006) 56−64.

[28] J. Brotzge, K. Droegemeier, D. McLaughlin, Collaborative Adaptive Sensing of the Atmosphere: New Radar System for Improving Analysis and Forecasting of Surface Weather Conditions. <http://dx.doi.org/10.3141/1948-16>, vol. 1948, pp. 145−151, 2007.

[29] OSTP (US Office of Science and Technology Policy), "Increasing Access to the Results of Federally Funded Scientific Research," 2013.

[30] C. Madurangi, I. Kouper, Y. Luo, I. Suriarachchi, and B. Plale, "SEAD 2.0 Multi-Repository Member Node," in DataONE Users Group meeting DUG-2016, 2016.