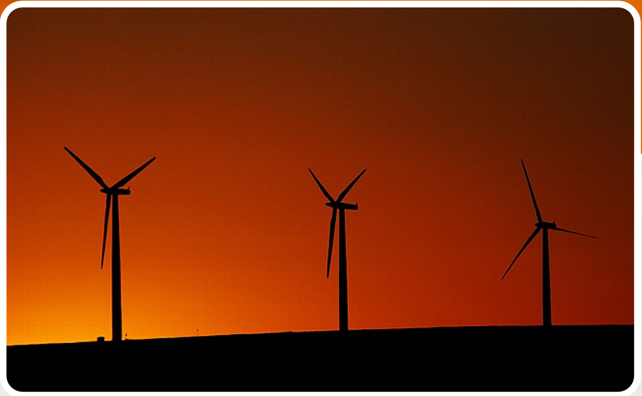
PNNL-23718

Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by **Battelle** *Since 1965*



Data Archive and Portal Thrust Area Strategy Report

September 2014

U.S. DEPARTMENT OF
ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, makes **any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights**. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
email: orders@ntis.gov <http://www.ntis.gov/about/form.aspx>
Online ordering: http://www.ntis.gov

This document was printed on recycled paper.
(8/2010)

PNNL-23718

Data Archive and Portal Thrust Area Strategy Report

September 2014

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99352

Executive Summary

The Data Archive and Portal (DAP) is a key capability of the U.S. Department of Energy's Atmosphere to Electron (A2e) initiative. The DAP Thrust Area Planning Group was organized to develop a plan for deploying this capability. The group consists of participants from Pacific Northwest National Laboratory, Argonne National Laboratory, Lawrence Livermore National Laboratory, the National Renewable Energy Laboratory, Sandia National Laboratories, the National Oceanic and Atmospheric Administration, the University of Michigan, Purdue University, and the University of Chicago. Additional input was solicited from subject matter experts in areas pertinent to the effort.

The DAP Thrust Area Planning Group envisions a distributed system—a DOE Wind Cloud—that functions as a repository for all A2e data. This Wind Cloud will be accessible via an open and easy-to-navigate user interface that facilitates community data access, interaction, and collaboration. DAP management will work with the community, industry, and international standards bodies to develop standards for wind data and to capture important characteristics of all data in the Wind Cloud. Security will be provided to facilitate storage of proprietary data alongside publicly accessible data in the Wind Cloud, and the capability to generate anonymized data will be provided to facilitate using private data by non-privileged users (when appropriate). Finally, limited computing capabilities will be provided to facilitate co-located data analysis, validation, and generation of derived products in support of A2e science.

There are a number of platforms outside of the Wind program that can serve as examples for managing the data. The planning group notes that no existing platform provides a ready-made solution for all of the A2e DAP needs. Instead, this document highlights the salient properties of five significant platforms that should be further considered and leveraged during deployment of the Wind Cloud.

The planning group recommends an architecture that comprises components to collect, preserve, process, discover, and access data. This architecture can be extended to create multiple instances of the DAP. It provides the capability to work with proprietary data that requires secure access and transmission or to accommodate existing resources. These instances would work together in an integrated fashion known as the Wind Cloud as described in this document.

Acknowledgments

The *Data Archive and Portal Thrust Area Strategy Report* was made possible based on the cooperative input from the DAP Thrust Area Planning Group members:

- Tom Boden, Oak Ridge National Laboratory
- Ann Brennan, National Renewable Energy Laboratory
- Mike Hagengruber, Sandia National Laboratories
- James Myers, University of Michigan
- Alex Pothen, Purdue University
- Rob Ross, Argonne National Laboratory
- Chitra Sivaraman, Pacific Northwest National Laboratory
- Dean Williams, Lawrence Livermore National Laboratory
- Glenn Rutledge, National Oceanic and Atmospheric Administration–National Climatic Data Center

The planning group also would like to thank Eric Stephan, Matt Macduff, and Clay Hagler, of Pacific Northwest National Laboratory, for their help with the technical solutions and architecture.

Acronyms and Abbreviations

A2e	Atmosphere to Electrons Initiative
ARM	Atmospheric Radiation Measurement Climate Research Facility
DAP	Data Archive and Portal
DOE	U.S. Department of Energy
ESGF	Earth System Grid Federation
PNNL	Pacific Northwest National Laboratory
SEAD	Sustainable Environment Actionable Data
NOAA	National Oceanic and Atmospheric Administration
OpenEi	Open Energy Information
UI	user interface
UQ	uncertainty quantification

Contents

Figures

1.0 Data Archive and Portal Objective

The Data Archive and Portal's (DAP) objective is to provide secure, timely, easy, and open access to all laboratory, field, and benchmark model data produced by the Atmosphere to Electrons Initiative (A2e) Initiative. This effort will provide interoperability among independently developed data to enable model runs and community interaction.

The DAP will collect, store, catalog, process, preserve, and disseminate all significant A2e data with state-of-the-art technology while conforming to or helping define industry data standards.

2.0 Thrust Area Objective

The DAP thrust area will build a U.S. Department of Energy (DOE) Wind Cloud (Mell and Grance 2011) to function as the single hub for all A2e data and, ultimately, all historical wind data. The Wind Cloud will take advantage of distributed systems and existing DOE-funded assets to solve computational problems and provide a single archive access point to distributed storage.

2.1 Goal 1: Collect, Store, and Preserve A2e Data

The DAP will function as the repository for all A2e data and ultimately all wind data. All facilities involved in creating wind data will have the ability to store that data in the Wind Cloud. The DAP also will provide sufficient, common computing infrastructure to support a range of services, hardware, and storage capabilities to facilitate collecting, storing, and preserving data.

2.2 Goal 2: Securely Store Proprietary Data

Security will be provided at multiple levels to allow for managing proprietary, embargoed, and full public access as appropriate. One option is to consider all incoming data as proprietary and provide secure transmission and access points.

2.3 Goal 3: Timely, Open, and Easy Data Access

A user interface (UI) will be developed to permit easy navigation for analysts and scientists. The UI will allow users to identify relevant data sets, understand content, capture provenance, and access subsets of interests. In addition, scientists will be able to access data relevant to their needs and in formats that enable efficient reuse and online display. The UI also will afford community interaction and collaboration.

2.4 Goal 4: Establish Industry Standards to Enable Data Discovery and Integration

A data catalog will be established for the wind community based on industry, community, and international standards. Persistent identifiers will be assigned to data. Metadata will be generated for all data regarding the data store, source, facility responsible, description, file format, database repository, attributes, security access, etc. To facilitate collaboration, underlying infrastructure will be developed to enable third-party tools and software to query and access the data.

2.5 Goal 5: Enable Science to Reduce Uncertainties in Models

Computing facilities and data test beds will be provided to rapidly improve and validate models and reduce uncertainties in models (as budgets permit). Data co-located with compute facilities and data integration tools will accelerate easy validation and promote data analysis. Software will be developed to produce best estimate derived products that can then be used as inputs to validate models.

3.0 User and Stakeholder Needs

To gather the requirements to support the A2e's Strategic Plan, a facilitated workshop was held with participants from several DOE national laboratories, universities, and institutions, including:

- Argonne National Laboratory
- Lawrence Livermore National Laboratory
- Pacific Northwest National Laboratory (PNNL)
- National Renewable Energy Laboratory
- Sandia National Laboratories
- National Oceanic and Atmospheric Administration (NOAA)
- University of Michigan
- Purdue University
- University of Chicago.

Subject matter experts were invited to present areas pertinent to this effort, and participants were asked to note gaps, challenges, or opportunities for each area as presentations where given. After each presentation, a conversation was facilitated and salient information captured.

One of the temptations or challenges often faced with requirements gathering is to jump toward technology or provide solutions. In an effort to mitigate these issues, a session was facilitated following these presentations using a technique called "brain writing" to capture and cull the requirements together (Wilson 2009). These functional requirements were gathered using the framework (Appendix A) with the goal that top-priority items are fleshed out into user stories and use cases after the work session.
At the end of the workshop, the user requirements were prioritized based on ease and impact (as described in Appendix A). The top-25 priorities (Appendix A) fell under the following categories:

- Ease of access/user friendliness
- Scalability
- Security
- Discoverability.

4.0 Existing Technologies

The DAP team reviewed several data management activities within the current Wind Program, as well as several platforms in use within the broader research community. Platforms outside of the Wind Program provide models that can be adopted or adapted by A2e. In the review, the planning group focused particularly on those funded by DOE that can be leveraged for A2e requirements. Each platform has been developed for a particular user community, a function within the software stack, or a programmatic goal. While no one platform provides a ready-made solution for all of A2e, together they represent a comprehensive toolkit that A2e program managers and researchers can deploy to meet requirements for data archival and access.

The platforms that were reviewed are described as follows.

Atmospheric Radiation Measurement Climate Research Facility (arm.gov), known as ARM, is a user facility that provides data from strategically located remote sensing observatories around the world. ARM offers data collected from instruments owned by ARM and for data collected by some agencies pertinent to ARM for the last 20 years. ARM maintains a fully integrated data system, from data collection to archival to dissemination. ARM also provides computing facilities and data visualization and analysis tools. Several DOE institutions and universities work together to develop a robust environment to collect, transfer, process, reprocess, and archive quality-controlled data. ARM also has established its own set of metadata standards to enable easy discovery and integration of data through their discovery web interface.

Earth System Grid Federation (esgf.org), or ESGF, is an international peer-to-peer network of more than 70 autonomous nodes for data storage and computation. Used primarily within the climate research community, the federation protocols provide single sign-on for users and allow each node to establish appropriate permissions for data access. Computational resources and visualization toolkits also can be accessed by users. ESGF supports several data transfer systems, including Globus. The software for establishing a node within the federation is open source. As a network of distributed repositories, ESGF could help A2e design the interactions and governance of the distributed network, including managing access and authentication. The software stack installed for each node in the system is modular and allows each node to offer different types of services, such as metadata registry, data delivery, and compute.

Globus (globus.org) is a web-based system for exchanging large scientific data sets among researchers and computational centers. Globus is well established within the DOE lab community as an efficient way to exchange very large data sets and model results that are stored locally where they are generated. The service for managing permissions and access is cloud-based. In 2014, Globus also is developing new data publishing and discovery capabilities that will allow data producers to publish metadata and permanent identifiers for their data sets and create targeted data collections (for example, for A2e) along with providing search capabilities for data collections and the network as a whole.

OpenEI (openei.org), or Open Energy Information, is a Wiki-style, cloud-based platform for public data that allows registered users to contribute content, data sets (in any format), images, and other energy-related information. It includes a community section for public or controlled-group interactions. OpenEI serves as a metadata catalog for data sets stored in other repositories, as well as direct storage and access for “smaller” data uploaded to OpenEI (currently limited to 1 GB per upload).

SEAD (sead-data.net), or Sustainable Environment Actionable Data, is sponsored by the National Science Foundation and addresses the data management and sharing requirements of the sustainability science field. SEAD is a relative newcomer in the data platform world, providing a number of capabilities that may be relevant for A2e. SEAD can manage any file type and any metadata vocabularies needed, and it allows continued annotation of data over time. It also focuses on active data use and reuse with features such as group “project spaces” for organizing data; the ability to import data from other sources as part of new collections; and the incorporation of a researcher profile service that maintains live links between people, published data, papers, projects, and organizations. SEAD also can package large numbers of files with rich metadata into aggregate standards-based research objects, simplifying long-term storage and discovery.

5.0 Data Flow

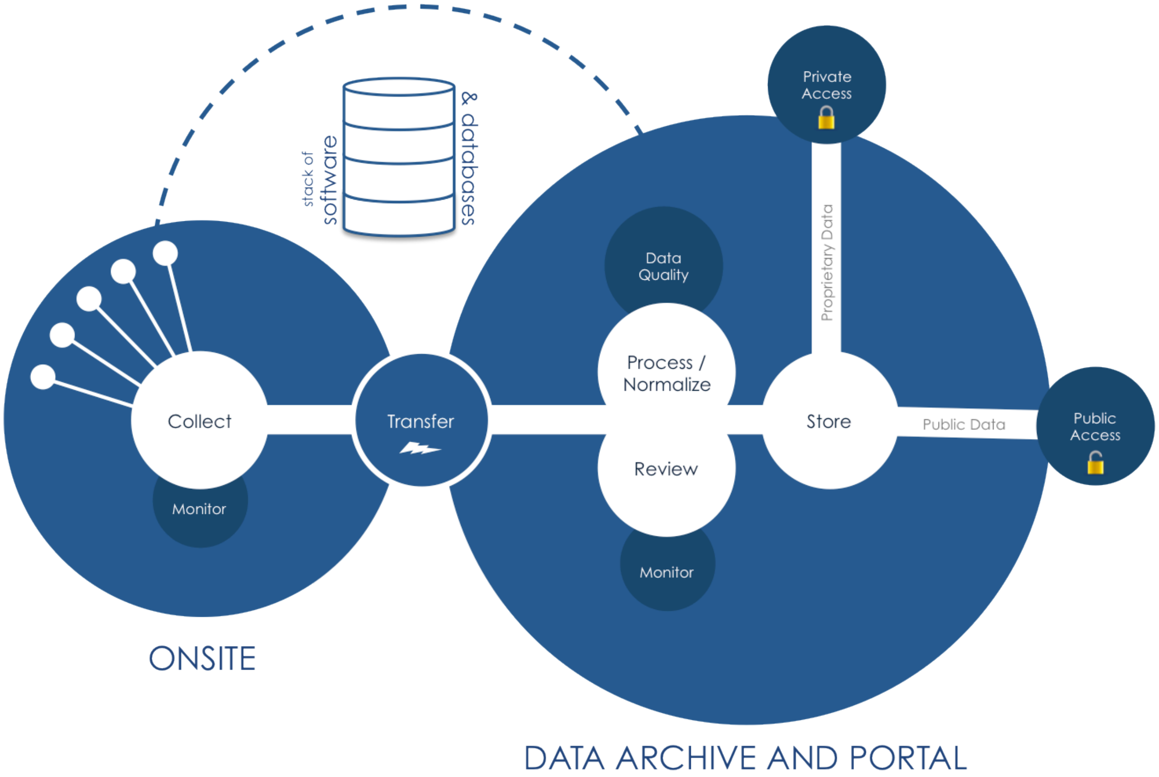


Figure 1. A model of the projected DAP data flow.

As projects produce data, DAP's goal is to collect, transfer, review, and archive the data and ultimately disseminate data based on established policies. The DAP will be responsible for building the infrastructure, including: software stacks and databases; standards, policies, and procedures to collect the data securely; transferring the data in a timely fashion; reviewing and monitoring the data; storing the data; and disseminating data through a user-friendly interface.

6.0 The DAP Architecture

The DAP architecture has been developed to support A2e DAP baseline requirements based on presentations provided by the subject matter experts and identified through interactions with the community and A2e thrust area leaders. The baseline requirements include:

- Collection and data preservation services
- Access to preserved and historical data
- Cataloging in support of data discovery
- Processing services to transform data based on user requirements.

The recommended model includes the collection, preservation, access, catalog, and processing components. Some key features of this model include the need for high-speed data transfer nodes for moving data in and out of the DAP. The externally facing components will be segregated to mitigate security risks to the rest of the system. Although this model represents a traditional data archive solution, it could be implemented with either institutional or cloud resources as a centralized solution.

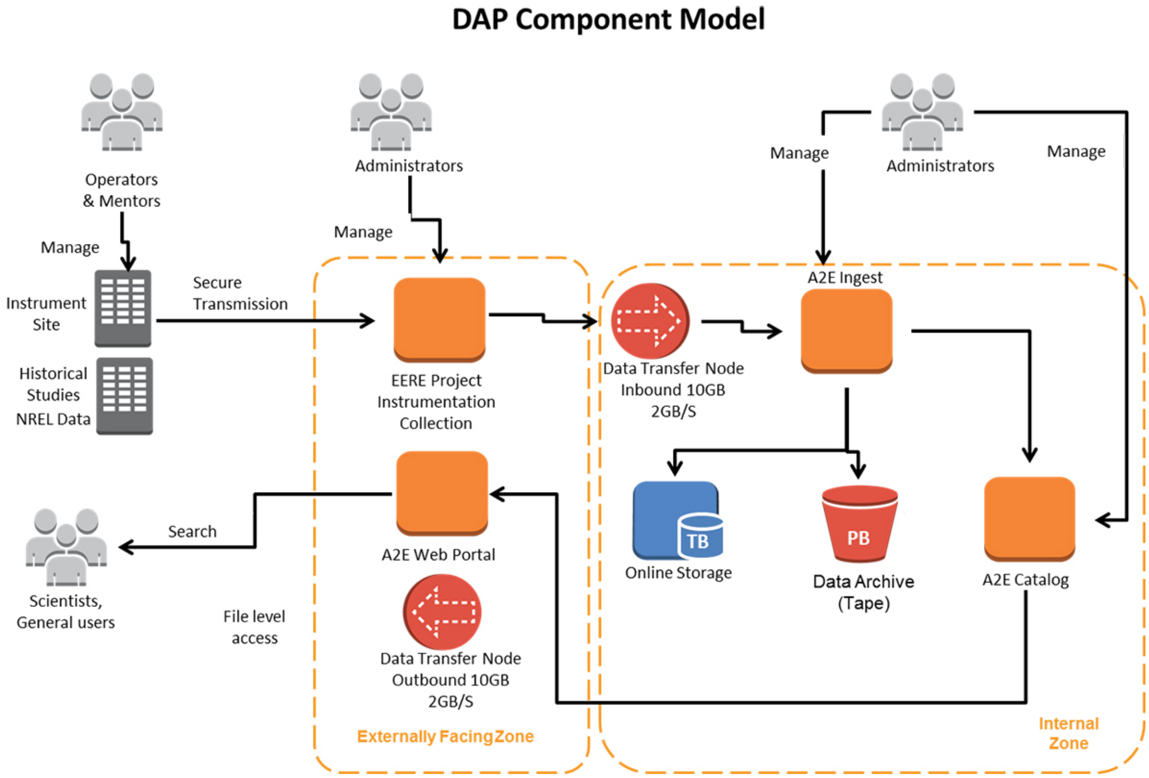


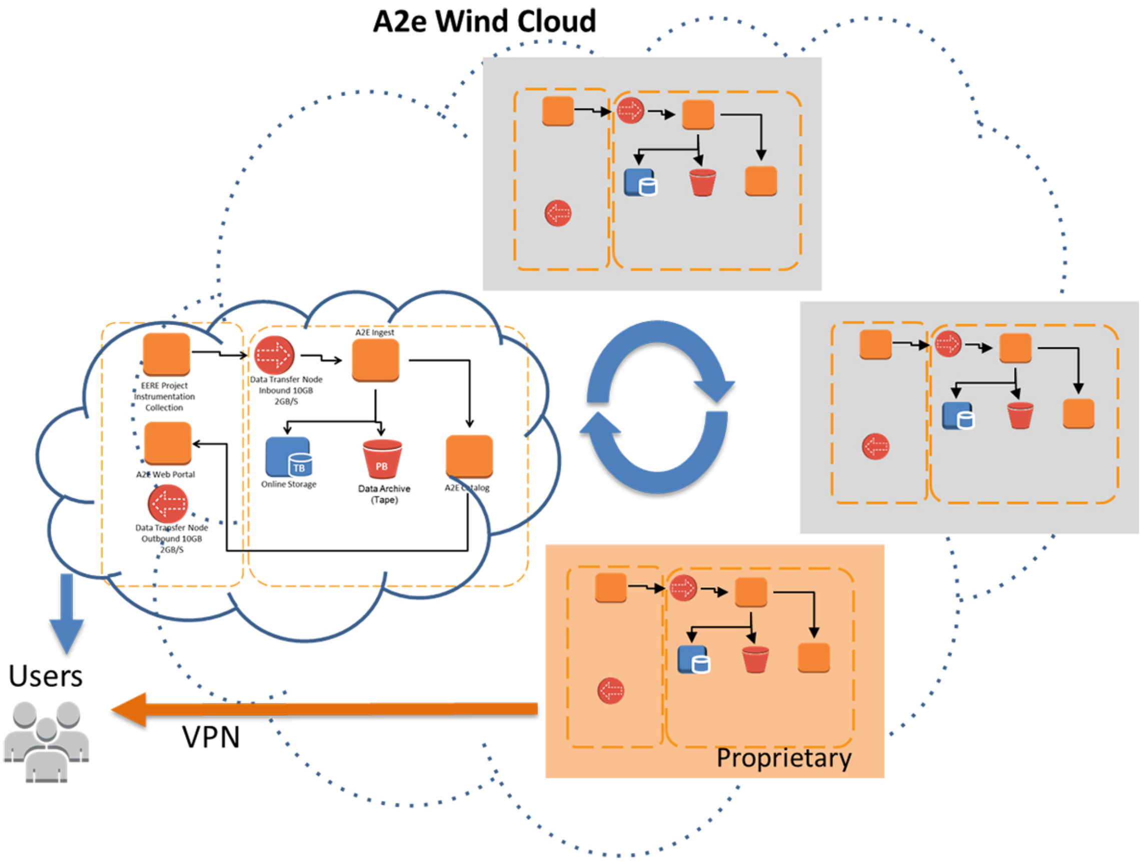
Figure 2. The DAP Component Model

The cost of building and maintaining this system will depend on the selected location and would vary depending on the data volumes and compute capacity. This model is scalable and realistic to deploy at a single location and can be extended to include other institutions.

6.1 A2e Wind Cloud

The realities of existing distributed institutional resources and the need for some data segregation are not fully realized by the DAP architecture alone. DOE may have access to existing institutional resources, such as compute and storage facilities that can be leveraged for the DAP. Because of unique computing and data use characteristics, the DAP model can be replicated at separate resources. The DAP model will include the ability to access data and metadata between instances. However, this ability would be restricted in the case of proprietary data stores.

For proprietary data sets, additional security measures and virtual private networks (VPNs) would be employed to protect the data. There will be increased management cost in creating and coordinating multiple DAP instances, which will be considered as part of Wind Cloud's overall growth. By adopting the DAP model and with the ability to replicate it as needed by the program, the Wind Cloud provides a solution that addresses the requirements and realities of data use. The advantages of adopting the DAP model are software reuse, meeting DAP metadata standards, preservation, and dissemination of data that enables single sign-on access.



A2e Wind Cloud
EERE Project Instrumentation Collection
Data Transfer Node Inbound 10GB 2GB/S
A2E Ingest
A2E Web Portal
Online Storage
TB
Data Archive (Tape)
PB
A2E Catalog
Data Transfer Node Outbound 10GB 2GB/S
Users
VPN
Proprietary

Figure 3. A2e Wind Cloud Architecture

7.0 Establishing Standards

Reusability of scientific data depends not only on its raw preservation and availability but also on users' ability to understand it. For digital data, this includes understanding the data format—how scientific information has been encoded as bits and bytes—and having information about how the data was produced and what it means (often called *metadata*). As data volumes have grown, there has been increasing recognition that formal documentation and/or standardization of data formats and metadata vocabularies are critical tools for effective data reuse at scale. Furthermore, to support integration of data from multiple sources and tracking data provenance (a type of metadata describing how data was generated) as data moves, it becomes critical to provide persistent data identifiers and standard mechanisms to retrieve data and metadata given an identifier. It also becomes important to standardize how large amounts of data are organized into collections and sub-collections.

With the diversity of data to be managed within A2e, it is unlikely that a single data format, or single metadata standard, will cover all needs. Indeed, it is becoming common for repositories to support multiple formats and vocabularies. Determining which formats and vocabularies to support becomes a matter of both policy and technology choices. For A2e, it is clear that self-describing data formats, such as NetCDF, will be important, but popular formats for storing spreadsheets, images, and videos, as well as formats related to specific data types such as LiDAR, also will be relevant. The A2e DAP will need to develop policies and procedures related to accepted formats or, more flexibly, how well different formats will be supported (e.g., if it will be possible to request subsets of data versus simply retrieving files).

There are a number of persistent identifier mechanisms relevant to A2e, such as Digital Object Identifiers (DOIs) that are gaining broad acceptance and Archival Resource Keys (ARKs) that simplify managing data in federated systems. Similarly, there are a range of vocabularies in common use for describing scientific data, ranging from the near-ubiquitous Dublin Core standard for basic bibliographic metadata to widely used standards for geospatial location information (Federal Geographic Data Committee) and provenance (W3C PROV) to emerging vocabularies for describing scientific collections and domain information, including variables measured, units, procedures, instruments, and models. The Data Catalog (DCAT) vocabulary is another relevant standard, supporting exchange of data set descriptions between catalogs, that builds upon several underlying vocabularies.

The A2e DAP must define a core set of required metadata, captured in these vocabularies, as well as allow (store and serve) any additional metadata that data providers believe is useful. The former ensures a useful minimal capability to find and use data, while the latter enables specific sub-communities to drive the state of the art as required.

Best practices in data preservation are rapidly evolving. The goal of providing basic reference data to a broad community is being replaced by the more ambitious goal of enabling reproducible research. This, in parallel with the dramatic growth in data volume, variety, and velocity, is driving an integration of technologies, processes, and procedures from distributed computing, library science, software engineering, and other domains.

To meet current and future demands, it will be important for A2e DAP to adopt a flexible design and manage scope through policies and procedures concerning what types of data to accept, how long different types of data will be stored, how formally metadata will be defined and reviewed, etc. For example, such an approach would allow A2e DAP to require that unique observational data be submitted in specific formats with full metadata documentation while also supporting storage of derived data sets in other formats with less formal review and shorter guarantees for how long the data would remain available. With such a flexible architecture, decisions about what to preserve (accession) and how long to keep it (retention) can be matters of policy and may change as needs evolve.

It is quite clear that the A2e DAP will need to evolve over the next decade. With increasing science automation and progressively more rich metadata, even the most basic assumptions about how data will be discovered and whether humans or computers will be the primary DAP users will need to be questioned. Vocabularies surely will change in the next decade. Notably, some of those mentioned herein only appeared within the last few years. As more is understood about how data preservation enables new research and as storage costs decrease, accession and retention policies will need to adjust. Similarly, increasing re-use may change community choices related to data formats, as well as the scale, scope, and purpose of collections. In other areas, research is being reorganized from storage-based on technique or discipline to site or problem-centric organizations. Furthermore, derived data, e.g., data that has been quality checked, recalibrated, refined, or otherwise improved through coherent integration of multiple source data sets, is gaining increased visibility and value. With a modular architecture and by using standard identifiers, vocabularies, and formats, the A2e DAP will be able to provide useful services to researchers today and grow as their needs evolve over time.

7.1 DAP Data Anonymization and Synthetic Data Generation

For Cooperative Research and Development Agreement (CRADA)-based projects and those with non-disclosure agreements, there will be a need to make a portion of the confidential or proprietary data streams available only to a subset of the public that meet the necessary access criteria. However, there also may be a need to share a version of the collected data with the general academic and research communities to further benefit the DOE Office of Energy Efficiency and Renewable Energy (EERE) missions. These seemingly conflicting goals can be accomplished by data anonymization and synthetic data generation. Figure 4 illustrates a sample workflow.

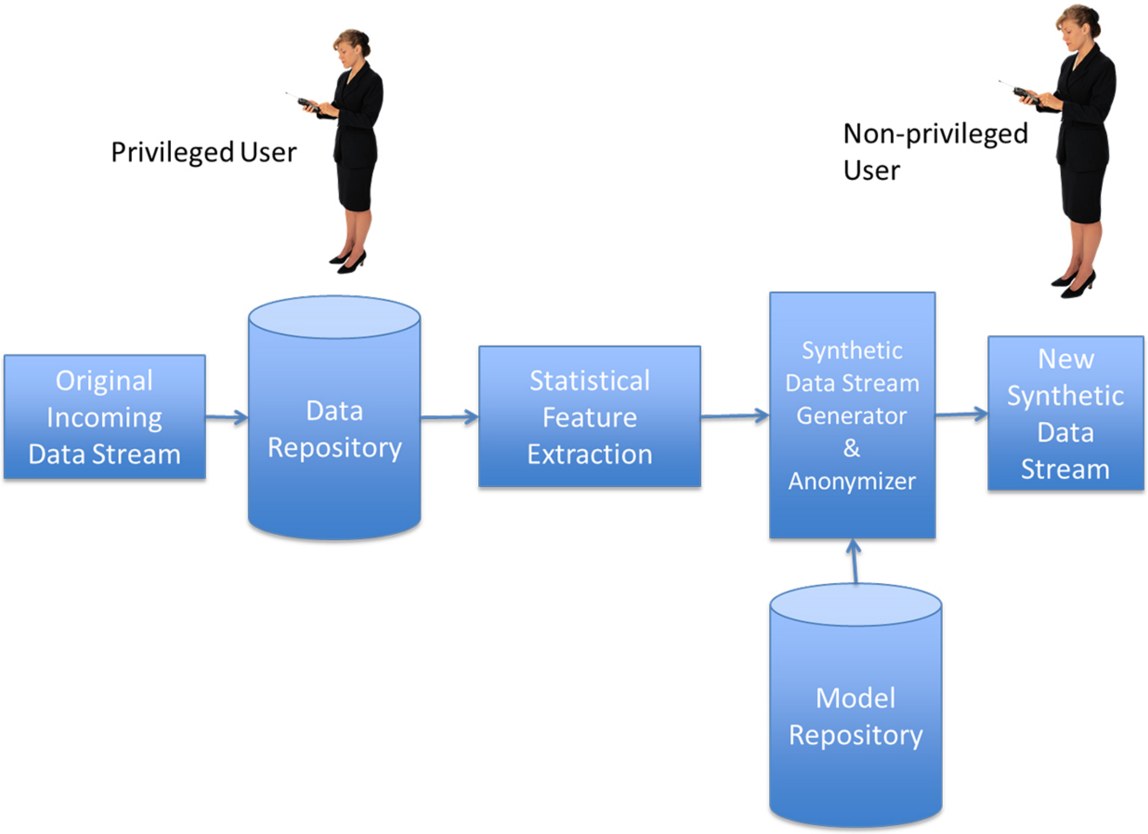

Privileged User
Non-privileged User
Original Incoming Data Stream
Data Repository
Statistical Feature Extraction
Synthetic Data Stream Generator & Anonymizer
New Synthetic Data Stream
Model Repository

Figure 4. The sample workflow shows how confidential or proprietary data streams can be made available to different users.

The original incoming data stream is saved into the repository, indicated by the normal workflow. This data is available to all users with access privileges to view the data stream. A process that runs on the repository periodically will inspect collected data as required and extract statistical features from the data stream based on a set of models that are constructed to fit the data. Once these features are extracted, a synthetic data stream generator will use the selected models and the extracted features to create a statistically similar and fully anonymized (e.g., no device names, locations, etc.) "synthetic" data stream for use by the general academic and research communities.

8.0 Work Packages

As described, the DAP architecture meets the key objectives of the DAP thrust area and is well aligned with the A2e's mission. Based on the preceding recommendations, the following work packages will be developed to track progress as the Wind Cloud is developed and deployed.

8.1 Work Package 1: Data Archive and Storage

A DOE Wind Cloud will be built using the DAP architecture developed in this document. It will act as the single hub for A2e data and ultimately all wind data. This cloud solution would provide a DOE repository for all wind data and single sign-on capabilities to access data that can be distributed. All facilities involved in storing wind data would have the ability to store and manipulate reference data inside the Wind Cloud.

All historical data will be stored, catalogued, and preserved at this archive, including all versions of the reprocessed data. The data repository will provide versioning and data replication for backup and better access.

8.2 Work Package 2: Security Management

A security system will be designed and developed to afford flexibility to have many types of access groups (from highly sensitive proprietary to publicly available data). The data owner also could specify the access level of the data when registering the data set in the catalog. Security could be designed to enable data access by roles, projects, and policies. A cyber security system also will be developed to manage portal security and secure data transmissions. Security management will not only exist for data access but for other resources, such as compute and storage resources.

8.3 Work Package 3: User and Service Interface

The front-end web interface will be designed, created, and integrated with a content management system for greater control over projects and project management content. To distinguish between the many projects in A2e, this web interface will combine the content management system, scientific online collaboration environment, interfaces to data and metadata services, and facilities for formal project governance. The web front end will be specifically designed for multi-project distributed organizations and integrated with back-end data services.

Tools will be an integral part of the interface's compute component and dissemination software stack. Tools will be created or leveraged from the community for remote processing and visualization within the enterprise system. Tools also will be made available to include model metrics, diagnostics, uncertainty quantification (UQ), comparative visualization, statistical analyses, etc.

In addition, the portal will include community collaboration features, as well as the ability to develop and publish best practices.

8.4 Work Package 4: Content Management and Standards

State-of-the-art systems will be built to support arbitrary data types, metadata, and processing workflows and allow imposition of file format restrictions, minimal metadata requirements, and curation and preservation processing to be managed through policy and configuration control.

State-of-the-art systems also will be built for parallelization of data access, file resolution (finding the physical location of data bytes give an identifier), and discovery queries (finding identifier(s) for relevant data based on metadata), leading to near-linear scaling with resources.

Standards will be developed for data, formats, control variables, metadata, and conventions for A2e. A governance committee will be created to control the process of creating and extending standards. The governance committee will be responsible for mandating and managing community needs and making decisions.

Resources will be dedicated to developing and maintaining ontologies and organizing data in a manner appropriate to enable the science activities.

8.5 Work Package 5: Algorithm Development, Processing, and Computing

Ingest processes will be developed to parse and structure raw files into standard formats. This process will ensure that minimum quality checks and calibrations are applied, and the data are converted to engineering units.

There always will be quantities of interest that are either impractical or impossible to measure directly or routinely. Established algorithms will be developed and implemented as value-added products, or VAPs, and will help fill some of the unmet measurement needs or improve the quality of existing measurements.

A testing-to-production environment for observation, simulation, and evaluation (with model metrics, diagnosis, UQ, and inter-comparison) also will be developed.

8.6 Work Package 6: Management

A governance committee will be formed to ensure that the DAP continues to meet user community and stakeholder requirements. This committee also will regulate and prioritize all tasks related to data products, computing and network infrastructures, processing architectures, software services, and tools. A work breakdown structure will be used to define, assign, track, and report tasks. Smaller committees would be formed to review the data and metadata and prioritize data migration to new formats and standards.

9.0 Integration with Other Thrust Areas

The DAP will interact with the experimental measurements, high-fidelity modeling, controls, and reliability thrust areas to collect, store, and preserve data generated by each areas. The DAP will act as a hub to strongly accelerate and advance model development, testing cycles, execution of high-fidelity models, and enable collaboration by automating labor-intensive tasks, providing intelligent support for complex tasks, and reducing duplication of effort. The DAP will enable discovery, as well as simulate and validate models. The DAP also envisions interacting with the integrated wind design plant thrust area to develop and improve reduced order models that will be used in the systems optimization during the design process. In anticipation of that effort, a survey was taken with the leaders of the other thrust area to understand their needs and requirements, and the DAP will continue to engage other thrust area teams to meet their requirements.

10.0 References

Wilson C. 2009. *Brainstorming and Beyond: A User-Centered Design Method*. Morgan Kaufmann Publishers, Burlington, Massachusetts.

Mell P and T Grance. 2011. *The NIST Definition of Cloud Computing*. NIST Special Publication 800-145. National Institute of Standards and Technology, Gaithersburg, Maryland. Available online: http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf.

Note: This publication identifies "five essential characteristics" of cloud computing:

1. *On-demand self-service*. A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.
2. *Broad network access*. Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations).
3. *Resource pooling*. The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model with different physical and virtual resources dynamically assigned and reassigned according to consumer demand.
4. *Rapid elasticity*. Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear unlimited and can be appropriated in any quantity at any time.
5. *Measured service*. Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

Appendix A

Appendix A

During the workshop, a work session was conducted to gather and prioritize the requirements. The participants were asked to answer the following question:

The future DOE Wind Cloud should _______ in a way that ____, so that _______.

This framework separates the requirements into the "What," "How," and "Outcomes" that participants feel should be covered to declare the effort successful. These are technology-agnostic requirements describing what and how the system should behave. Then, these requirements can be used for solution selection and architecture design to ensure solutions meet or exceed the program's needs. These requirements were categorized and prioritized based on ease to implement and impact to the program. The top-25 requirements then were used as baseline requirements for the DAP.

Table A.1. Top 25 Wind Cloud System Requirements (Session Results)

Item	Category	Users	Ease	Impact
File transfer protocol	Accessibility	Researcher	5	5
Handles episodic data and continually streaming	Data flow: real-time/intensive operation periods	Producer	5	5
Monitor data access	Metrics	DOE	5	5
Dashboard	Metrics	DOE	5	5
Make public data truly public and useable	Open	Researcher	5	5
Accession & retention policies	Policies	Industry	5	5
Store any/all types of data	Scalability	Producer	5	5
Protect proprietary data and avoid Freedom of Information Act release (unintended)	Security	Industry	5	5
Rely on data existing standards (community) vocabularies	Standards	Researcher	5	5
Present a low technical barrier of entry for smaller studies	User friendly	Researcher	5	5
Publish data		DOE	5	5
Notification	Replication		5	5
Scriptable access protocols	Accessibility	Researcher	4	5
Restful front end to make fast	Accessibility	Researcher	4	5
Metadata and data should be coupled	Catalog	Researcher	4	5

Item	Category	Users	Ease	Impact
Federated master catalog	Catalog	Researcher	4	5
Categorize data by user needs	Catalog	Researcher	3.5	5
Allow data to be searchable and discoverable	Searchable/Discoverable	Researcher	3.5	5
Secure data	Security	Industry	3.5	5
Access control by data set	Accessibility	Industry	4	4
Comma-separated values formats during dissemination	User friendly	Researcher	4	4
Be modern user experience (UX)	Accessibility	Researcher	3	5
Leverage existing resources to save money	Leverage	DOE	3	5
Scale in terms of capability (numbers files/deliverables, users, amount/types of metadata)	Scalability	Researcher	3	5
Standardized formats	Standards	Researcher	3	5

Table A.2. Wind Cloud System Requirements (Additional Suggestions)

Item	Category	Users	Ease	Impact
Easily accessible (outside firewall)	User friendly	Researcher	3	5
Versioning	Replication		3	5
Single sign-on access	Accessibility	Researcher	3	4
Provide ad hoc and temporal data sharing	Community interaction	Researcher	4	3
Copies at multiple locations	Preservation	DOE	3	4
Provenance	Provenance		4	3
Allow new requirements and opportunities to evolve over time	Scalability	Researcher	3	4
Computing hardware	Computing facilities	Researcher	2	5
Well documented open DAP open archival systems/tutorials	Open	Researcher/DOE	2	5
Provide Data Quality reports	Quality	Researcher/DOE	2	5
Store and preserve important historical data	Preservation	DOE	2	4
Aggregate data before dissemination	Aggregation	Researcher	2.5	3

Item	Category	Users	Ease	Impact
Support user computing (computing not necessarily part of DAP)	Computing facilities	Modeler	1.5	5
Compute (derive data)	Computing facilities	Researcher	1.5	5
Analysis and visualization	Computing facilities	Researcher	1.5	4
Be part of an ecosystem of data tools	Data tools	Researcher	1.5	4
Manage info about non-data science objects, e.g., instruments, codes, sensors, projects, people	Provenance		2	3
Increase research reproducibility		DOE	2	3
Increase research repeatability		DOE	2	3
Should provide consulting services	Operations and outreach	Operations	1	5
Be interoperable with other cloud-based systems	Extensibility		2	2
Provide for standardized access to data with disparate raw structures and quantities	User friendly	Researcher	1	3
Data design catalog	Catalog	Researcher		
Provide for exchange and data to enable collaborative research to solve large computational problems	Community interaction	Researcher/DOE		
Allow coordination of work	Community interaction	Researcher/DOE		
Should allow researchers to focus on research	Community interaction	Researcher/DOE		
Have a way to ease people into (using) it	Ease	Researcher		
Store data/archive data	Preservation	DOE		
Preserve data for future research	Preservation	DOE		
Preserve data per Presidential mandate	Preserve	Researcher/DOE		
Preserve wind observation data	Preserve	DOE		
Replicate	Preserve	DOE		
Explain data origin	Provenance			
Data quality reports	Quality	Researcher/DOE		
Extensible APIs for continued development	Scalability			
Scalable POC > Regional > National	Scalability			

Item	Category	Users	Ease	Impact
Scalable for future large-scale data	Scalability			
Able to compete with external systems regarding usability	Scalability			
Secure data	Security	Industry		
Provides security and supports open data	Security	Industry		
Storage and provide secure access when needed	Security/preserve	Industry		
Have access methods that work for users	User friendly	Researcher		
Help researchers manage data during projects, i.e., to avoid duplicate management	User friendly	Researcher		
Balance strict requirements without frustrating users or pushing them away	User friendly	Researcher		
Capability to visualize data	Visualization	Researcher		
Support human and machine interactions				
Provide configurations facilities for modelers				
Be cost effective; maximize value/cost to A2e		DOE		
Accessible	Accessibility			
Advances sciences	Advance science			
Enables collaboration	Community interaction			
Cost efficient	Cost efficient			
Efficient	Efficient			
Easily integrated	Extensibility			
Flexibility for many applications	Extensible			
High quality	High quality			
Incremental	Incremental			
Intuitive	Intuitive			
Organized	Organized			
Reliable	Reliable			
Robust	Robust			

Item	Category	Users	Ease	Impact
Robust	Robust			
Scriptable	Scriptable			
Secure	Security			
Secure	Security			
Secure	Security			
Sustainable	Sustainable			
Sustainable	Sustainable			
Easy to use for data non-experts	User friendly			
User friendly	User friendly			
Get is used and adds value				
Dynamic self/service				
Cool				
Financially transcript $/Turbine				
Modeler for funding resources				
Lower cost of wind energy	A2e mission			
Advance wind technology	A2e mission			
Wind energy cost drops faster	A2e mission			
We know when where and how the wind blows	A2e mission			
High penetration of wind on grid	A2e mission			
Wind competes without subsidies	A2e mission			
Produce credible research findings	Advance science			
Science is done that could not be done without it	Advance science			
Reduce noise pollution from wind turbines	Advance science			
Better wind turbines	Advance science			
Create new A2e models	Advance science			
Workflow dataflow automate	Automate			

Item	Category	Users	Ease	Impact
Models and data are easily brought together in verification and validation (V&V) exercises	Community interaction			
Redundant data collection is avoided	Duplication of efforts			
Gaps in data are ready identified	Identify gaps			
It helps policy makers make decisions	identify gaps			
Federal investment is targeted to high impact	Identify gaps			
Inspire a new generation of energy researchers	Inspire			
DAP’s impact is clear (metrics)	Metrics			
Data can be repeated and reproduced	Reproducibility			
Secure data per the mandate	Secure data			
Other programs have a model/adaptable working capability				
We better integrate all wind data				
Less air pollution from fossil fuels				

Pacific Northwest
NATIONAL LABORATORY

U.S. DEPARTMENT OF
ENERGY

Proudly Operated by **Battelle** *Since 1965*

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99352
1-888-375-PNNL (7665)

www.pnnl.gov