International Journal of Geo-Information

MDPI

Article

A Semi-Automated Workflow Solution for Data Set Publication

Suresh Vannan *, Tammy W. Beaty, Robert B. Cook, Daine M. Wright, Ranjeet Devarakonda, Yaxing Wei, Les A. Hook and Benjamin F. McMurry

Environmental Sciences Division, Climate Change Science Institute (CCSI), Oak Ridge National Laboratory, Oak Ridge, TN 37831-6290, USA; beatytw@ornl.gov (T.W.B.); cookrb@ornl.gov (R.B.C.); wrightdm@ornl.gov (D.M.W.); devarakondar@ornl.gov (R.D.); weiy@ornl.gov (Y.W.); hookla@ornl.gov (L.A.H.); mcmurrybf@ornl.gov (B.F.M.)

* Correspondence: santhanavans@ornl.gov; Tel.: +1-865-241-6181

Academic Editors: Constanze Curdt, Christian Willmes, Georg Bareth and Wolfgang Kainz
Received: 20 December 2015; Accepted: 25 February 2016; Published: 8 March 2016

Abstract: To address the need for published data, considerable effort has gone into formalizing the process of data publication. From funding agencies to publishers, data publication has rapidly become a requirement. Digital Object Identifiers (DOI) and data citations have enhanced the integration and availability of data. The challenge facing data publishers now is to deal with the increased number of publishable data products and most importantly the difficulties of publishing diverse data products into an online archive. The Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC), a NASA-funded data center, faces these challenges as it deals with data products created by individual investigators. This paper summarizes the challenges of curating data and provides a summary of a workflow solution that ORNL DAAC researcher and technical staffs have created to deal with publication of the diverse data products. The workflow solution presented here is generic and can be applied to data from any scientific domain and data located at any data center.

Keywords: data ingest; publication workflow; data deluge; terrestrial ecology; automation; aeospatial

1. Introduction

Up until the early 1990s, terrestrial ecology data publication comprised primarily of graphs, tables, and figures included in published manuscripts. Data created during the process of research was often lost in the highly derived visual representations included in peer-reviewed literature. Furthermore, the data summarized in publications cannot be readily extracted for further analysis, let alone the integration of data into new research. Substantial effort is needed to convert the data into a format that makes the data usable. Since early 2000, research communities, funding agencies, and data users realized the need for publishing a "soft" copy version of the investigator-generated data.

A workflow for data ingest of the "soft" copy version of the data has been developed at the Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC) that is based on practices for data archival that formalizes interactions with users, compiles information, data files, and metadata, and releases the product to the public. This paper summarizes the challenges of curating data produced by the terrestrial ecology community and provides a workflow solution to efficiently archive diverse data products generated by researchers.

2. ORNL DAAC

The ORNL DAAC for biogeochemical dynamics is one of the National Aeronautics and Space Administration (NASA) Earth Observing System Data and Information System (EOSDIS) data centers managed by the Earth Science Data and Information System (ESDIS) Project [1,2]. The ORNL

DAAC (http://daac.ornl.gov) archives data produced by NASA's Terrestrial Ecology Program and as such provides data and information relevant to biogeochemical dynamics, ecological data, and environmental processes, critical for understanding the dynamics relating to the biological, geological, and chemical components of Earth's environment. The mission of the ORNL DAAC is to assemble, distribute, and provide data services for a comprehensive archive of terrestrial biogeochemistry and ecological dynamics observations and models to facilitate research, education, and decision-making in support of NASA's Earth Science. The data publication workflow system described here was developed in support of the ORNL DAAC mission to handle the diverse data products created by its user community.

3. Data Ingest—Essential 5Ps

The term ingest is used to refer to the process of building a data collection that involves capture, translation, organization, and registration of data [3] Ingest is a widely used term in a data archive and can be broken down into the following high-level tasks:

(1) Accepting the data package from the data providers, ensuring the full integrity of the transferred data files (through checksums, file counts *etc.*);
(2) Identifying and fixing data quality issues;
(3) Assembling detailed metadata and documentation, including file-level details, processing methodology, and characteristics of data files;
(4) Developing a discovery tool that allows users to search metadata for the data sets needed;
(5) Setting up data access mechanisms;
(6) Re-packaging data files to better suit the end user's research/application needs (optional);
(7) Setup of the data in data tools and services for improved data discovery and dissemination (optional);
(8) Registering the data set in online search and discovery catalogues;
(9) Provide a permanent identifier through Digital Object Identifiers (DOI).

Long-term storage, data stewardship, and user support are also considered while ingesting a data set into an archive. The nine tasks described here form the critical 5-Ps of data archive: Presentation, Preservation, Persistence, Publication, and Protection (5-Ps) are essential elements for digital repositories. For example, the Open Archival Information System (OAIS) framework that can be used as a reference model for repositories contains all of the 5-P elements [4]. A short summary of each of the 5-P elements is provided. A detailed description of the ORNL DAAC activities as it relates to these 5-P elements is available at the ORNL DAAC Data Management for Data Providers website [5].

Presentation: Archive tasks conducted for staging the data for end user access.
Preservation: Activities conducted at the archive to ensure continued access to the data for as long as necessary.
Persistence: Creation and use of persistent identfiers for the archived data products.
Publication: The task of moving the ingested data to publication. This includes creation of data citation, registration of DOI, activation of DOI data set landing page, registration of data set meta data, announcing data release.
Protection: The archived data are protected using various backup strategies (local and remote backup). Restricted data are also protected from public use using various access control mechanisms.

For proper data stewardship all the 5-P elements need to be considered and followed. In addition, the 5-Ps for data archive are generally consistent for all data archive projects, irrespective of science domain and diversity of the data products. The UK data archive lists the process for data archive and includes all the elements described above [6]. The National Oceanic and Atmospheric Administration (NOAA) National Centers for Environmental Informatcis (NCEI) data publication system also presents various tasks related to publication of a data set [7]. While the ingest tasks may be similar across various

archives, the execution of the task is much more labor intensive for publication of heterogeneous data products.

Efficient data management during all stages of a project is essential for effective data preservation and publication. In particular, policies and technical strategies for management of heterogeneous data are especially important. The 2005 National Science Foundation report on long-lived data collection illustrates the importance of policies and strategies to facilitate the management, preservation, and sharing of digital data. The report also expressed a need to fully embrace the essential heterogeneity in technical, scientific, and other features found across the spectrum of digital data collections [8]. Heterogeneous data are especially difficult to manage and ingest. Since 2005, considerable progress has been made on the policy aspect of data management. Several US research agencies have adopted a public data policy and are requiring a data management plan as part of a research project [9]. Some journal publishers are also requiring that data used/generated as part of a research publication be published and linked in the references prior to manuscript publication [10,11]. While these policies are excellent for data preservation and provenance, the technical ability of archives to handle the data deluge needs to be addressed especially for handling heterogeneous data sets. The workflow solution described here presents an informatics system that minimizes the labor related to tracking and ingest of heterogeneous data products into an archive while considering all the 5-P elements essential for good data stewardship and archive.

4. Data "Deluge" and Diversity—ORNL DAAC Case Study

The data created by the terrestrial ecology community is highly heterogeneous in nature. As of 31 January 2015, the ORNL DAAC had ingested 1126 terrestrial ecology data sets over 21 years (1994 to 2015) to its online archive. The average ingest time for a single data set is about 6 days. The ingest time period ranges from 2 days to several months in rare cases for complex data that requires extensive iteration with the data creators. In addition, the archive provides post-project user support and ensures the long-term integrity of the data files. Relevant data sets are further curated and added to web based data visualization, subsetting, and data access services.

The volume and complexity of available data sets for ingest has increased since 2008. The ORNL DAAC has doubled in archive size from ~90 TB in 2013 to ~183 TB in 2015. Figures 1–3 show the diverse nature of the 1000+ data sets ingested into the archive over the last 20 years. Figure 1 shows the number of files per data set. The range is 1–92,619 files per data set. Figure 2 shows the histogram of the data volumes in the ORNL DAAC data holdings for which the range is 1 KB to more than 10 GB for several data sets. Figure 3 shows the various file formats archived at the data center. As shown in Figure 3, Binary (*.dat) and compressed zip (*z,gz) files constitute 95% of the data files in the archive. Of this 55% of the files are compressed Geostationary Operational Environmental Satellite system (GOES) satellite data containing images in binary format. Binary format was a popular storage format in the 1990s when image file format standards were immature. These binary formats often require special viewers or code (often written in FORTRAN) for analysis. It should also be noted that ~33% of the ORNL DAAC holdings contain 95% of these binary data sets. The binary file format presents data curation challenges, but their use has diminished significantly since they were archived. The ORNL DAAC is actively curating and processing these binary compressed files into community standards such as netCDF and geotiff.

In addition to these data file diversity characteristics, here are some additional metadata characteristics of the ORNL DAAC archived data sets.

- 2583 keywords
- 1364 investigators
- 343 variables
- 282 sensors
- 125 sources (satellites, flux towers, airplanes, *etc.*)

Please note that the numbers represented above are as of 31 January 2015. With each new data set ingested into the archive the complexity of the data sets and corresponding metadata increases.

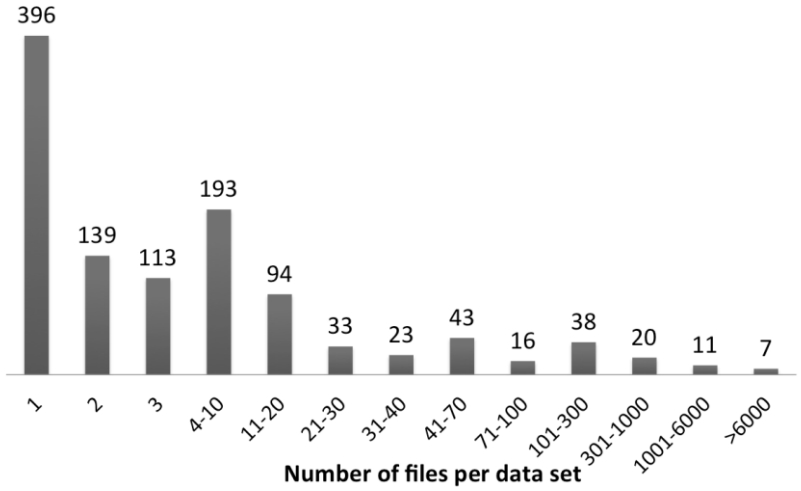

Figure 1. Distribution of number of files per data set in the Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC) data holdings (n = 1126).

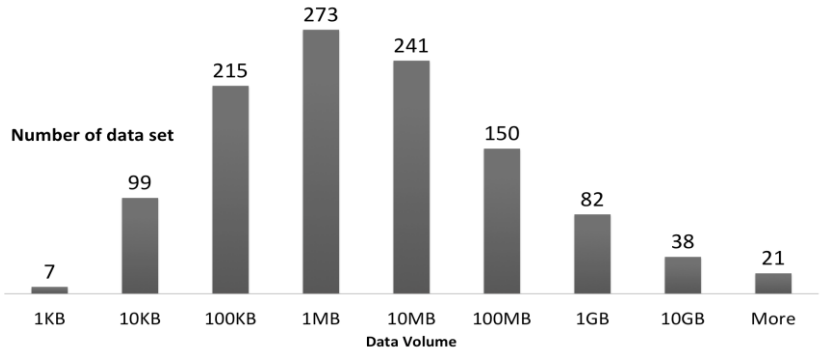

Figure 2. Data volume distribution of data sets in the ORNL DAAC data holdings (n = 1126).

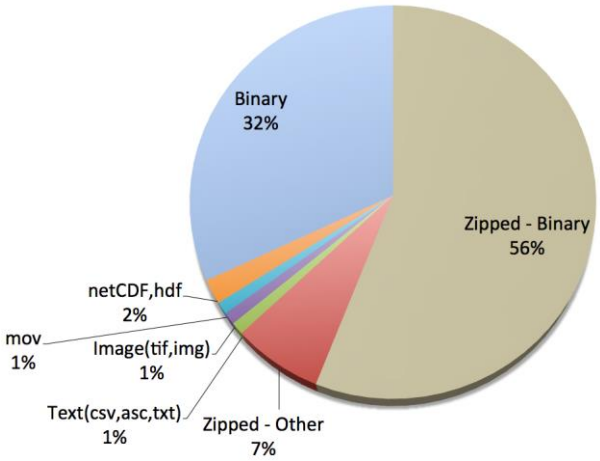

Figure 3. File format of ORNL DAAC data holdings (n = 1126).

The ingest rate, number of available data sets, and diversity of the data products presents challenges of curation with fixed work year effort. With the US government policy changes related

to the availability of data products, more data sets are potentially available for archive. To deal with these challenges and to improve efficiencies in the curation of heterogeneous data products, the ORNL DAAC has created a semi-automated ingest workflow system. Other data repositories face similar challenges and have developed similar data publication systems. For example, NOAA's Send2NCEI (S2N) online tool facilitates the publication of NOAA data files (oceanographic, atmospheric, and geophysical data) to the NCEI archive [7].

The following sections describes the ORNL DAAC **S**emi-**Au**tomated ingest **S**ystem **(SAuS)** and its benefits. The SAuS is a unique workflow solution developed at the ORNL DAAC for publishing heterogeneous terrestrial ecology data sets. DAAC staff used their collective experience with archiving terrestrial ecology data, heretofore a highly manual process, plus the practices of data curation to develop SAuS. Although the workflow has been developed to deal specifically with the terrestrial ecology data, the workflow concepts are applicable to data sets from other science domains.

5. Semi-Automated ingest System (SAuS)

The SAuS ingest workflow system was designed to streamline and improve the efficiency data ingest into the ORNL DAAC. SAuS is designed using the 5-P rule for effective data stewardship (Figure 4). The goal of the workflow system was to reduce the time taken to ingest a data set, to increase the quality of documentation and metadata, and to track individual data sets through the data curation process. A publication system modeled after a manuscript publication process is critical for handling data generated by individual investigators.

Figure 4. Semi-Automated ingest System(SAuS) design based on essential 5-Ps for effective data stewardship.

5.1. Workflow Architecture

Figure 5 shows a high level overview of the workflow architecture. The SAuS data publication workflow system can be broken into two broad categories, the data provider interaction process and the ORNL DAAC curation process. The data provider interaction consists of an inquiry for archive followed by the submission of the data set(s). The curation processes are the detailed steps followed by archive staff for quality analysis, preparation of metadata, assembly of data files, citation, and documentation preparation. If the information compiled during the first part requires clarification, archive staff will iterate with the data provider to address any additional questions.

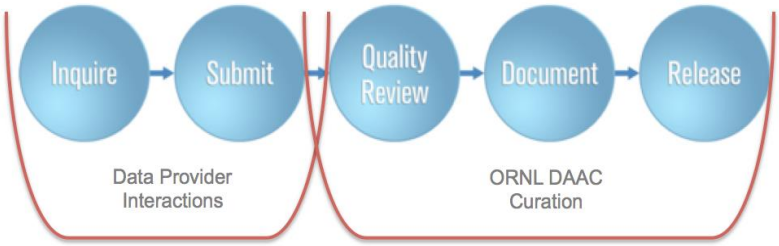

Figure 5. High-level representation of the SAuS workflow.

Figure 6 show the detailed steps involved in the SAuS workflow and the various ingest roles involved in the process. The boxes highlight the various tasks color-coded by the primary role. The Data Provider (DP) is the user submitting the data to the archive. The Ingest Coordinator (IC) is the person responsible for the interface between the data provider and the ORNL DAAC internal ingest activities. The IC oversees the DP submission, assignment of Quality Analysis (QA), and documentation, verification of the data set package, and the publication of the data set. Each data set is assigned a QA lead who has geospatial data analysis skills. The QA lead performs and data transformations and collection of granule level metadata. Each data set is also assigned a Documentation Lead (DL) . The DL prepares manuals describing the content of the data sets. The DL also collates any additional information collected during the QA process and integrates them into the documentation. Each data set is published only after it is approved by the ORNL DAAC chief scientist (DS). The DS verifies the scientific integrity of the documentation, data files, and ensures that the information added to the archive is relevant to the user community. Although the figure is presented somewhat linearly, there is a lot of communication and iteration between the various roles and stages within the ingest process. Further details of the individual steps in the workflow are described in the following sections.

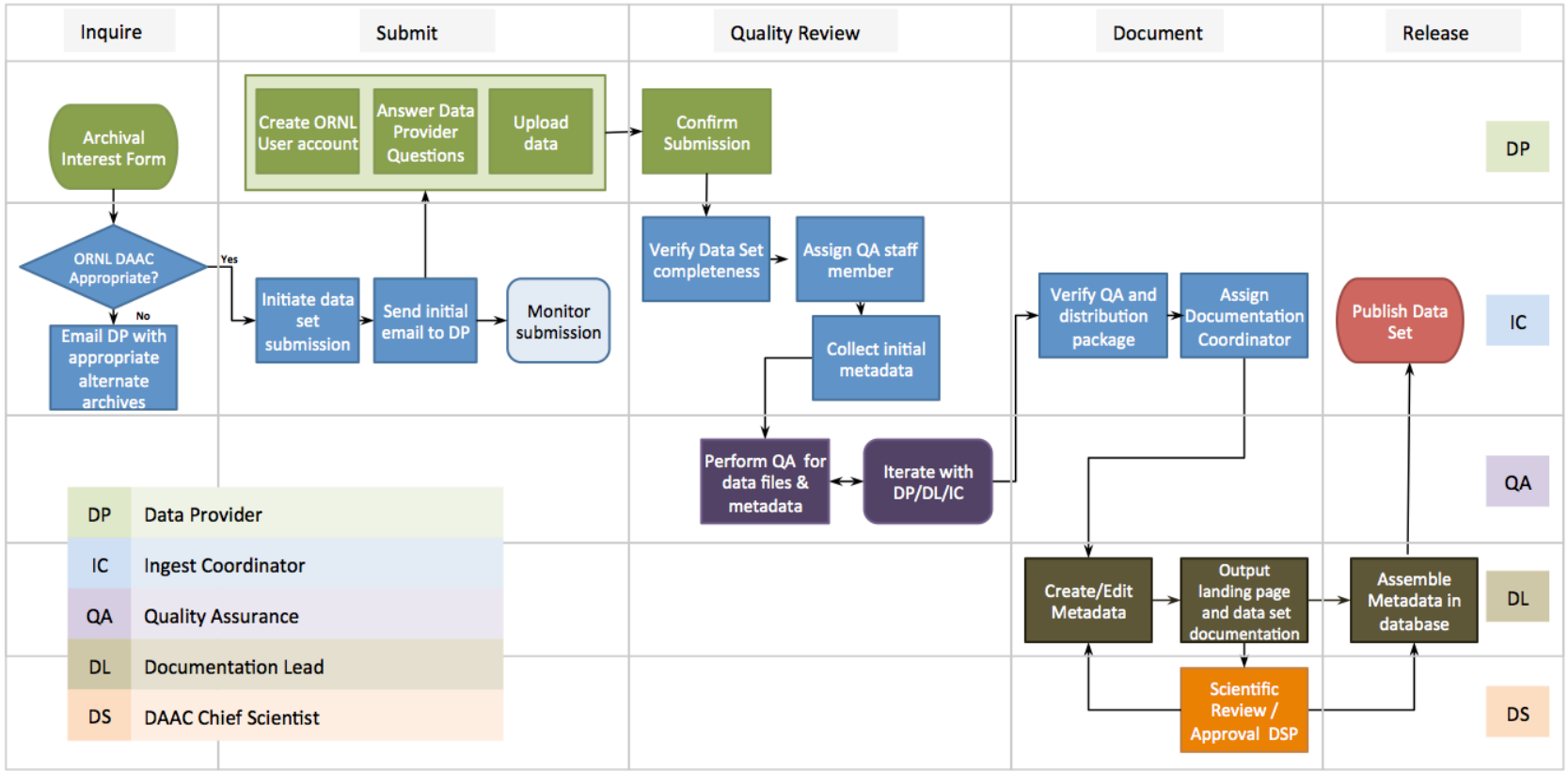

Figure 6. Detailed View of the SAuS workflow with various ingest roles defined (Inquire, Submit, Quality, Document, and Release).

5.2. Data Provider Interactions

Initial interactions with data providers occurs primarily through the ORNL DAAC's website on data management and data archival [5]. The website contains tutorials and training and reference material on data management [12,13]. In addition, the webpages provide an overview of data management planning and preparation and offer practical methods to successfully share and archive your data. Information on these pages serve as background and help files for the data activities associated with SAuS, including the steps providers need to take to initiate data ingest through this workflow.

Stage 1 (Inquire and Submit): The ORNL DAAC receives notification of a data provider's interest in archiving a data set through an archival interest form (Inquire). The form is available on the ORNL DAAC homepage and is prominently displayed. The *archival interest form* captures preliminary information about the data set, such as the funding source, data set title, data set description, *etc.* The details submitted by a potential data provider are added to a database at the ORNL DAAC archive center. When an *archival interest form* is submitted, the SAuS system triggers an email to the data coordinator and the ORNL DAAC chief scientist. A decision about archiving is made based on archive policy and data set priority, as approved and endorsed by EOSDIS and the ORNL DAAC User Working Group [14]. At the ORNL DAAC, data sets from the NASA Terrestrial Ecology and related programs (Carbon Monitoring Systems (CMS), Interdisciplinary Science (IDS), Carbon Cycle Science, *etc.*) are accepted and processed in the same order they are received. The order in which the data sets are processed may be adjusted based on the condition and quality of data and documentation when received and how quickly investigators respond to questions. If the candidate data sets do not fall within the purview of the ORNL DAAC archive, the user is notified and possible alternative data archives are suggested to the data provider. Non-NASA sponsored data sets that are directly relevant to the terrestrial ecology community have to be approved by EOSDIS and the User Working Group prior to archival, based on the importance of the data set for the community, data set size and condition, and resource availability.

Stage 2 (Quality Review, Document): If the data set is selected for archive, the data set moves into the "active" ingest phase. Through the SAuS publication dashboard, the ORNL DAAC ingest coordinator triggers an automated email to the data provider. The SAuS publication dashboard is a Drupal-based content management system that provides a graphical user interface for tracking and moving data sets through the ingest workflow. A description of the dashboard is provided later. When the automated email is triggered, automated scripts create various staging areas for the data sets. A unique data set ID based on a Universally Unique Identifier (UUID) is created for ingest in SAuS. The UUID is used to create a data upload directory and to add relevant information about the data set into the ingest workflow tracking database. An email address based on the UUID enables tracking of all email messages associated with the data set. The initial email to the investigator triggered through the dashboard contains four key elements.

(1) Information to get an user account on the ORNL DAAC data publication system;
(2) Link to answer a short questionnaire about the data set;
(3) Link to upload data files;
(4) Link to notify the ORNL DAAC system that all the above steps are complete.

The user account allows authenticated access to the data publication system and provides an added level of security to the ingest workflow. The short questionnaire gathers preliminary information about the data set to assist with quality assessment of the data files and to build data set documentation. The ORNL DAAC data set ingest questionnaire was designed from user community input and is aimed at maximizing the information (metadata) collected from the data provider in a reasonable amount of time (~30 min) to expedite the data publication process while retaining ingest quality. A summary of the questionnaire is provided in Table 1. In addition to answering the questions, the data provider submits the completed data or model products, including description documents

and supplemental files, using the UUID specific upload area. After the files and the answers have been uploaded, the data provider verifies the completion of the steps, which closes the ingest submission stage. The SAuS system provides mechanisms to send reminders and can snapshot the answers and the file upload summary information for provenance and record keeping.

Table 1. Data Provider questions [15].

Information About Your Data Set
Have you looked at our recommendations for the preparation of data files and documentation?
Who produced this data set?
What agency and program funded the project?
What awards funded this project? (comma separate multiple awards)
Data Set Description
Provide a title for your data set.
What type of data does your data set contain?
What does the data set describe?
What parameters did you measure, derive, or generate?
Have you analyzed the uncertainty in your data?
Briefly describe your uncertainty analysis.
Will the uncertainty estimates be included with your data set?
Temporal and Spatial Characteristics
What date range does the data cover? (YYYY-MM-DD)
What is a representative sampling frequency or temporal resolution for your data?
Where were the data collected/generated?
Which of the following best describes the spatial nature of your data? (single point, multiple points, transect, grid, polygon, n/a)
What is a representative spatial resolution for these data?
Provide a bounding box around your data.
Data Preparation and Delivery
What are the formats of your data files?
How many data files does your product contain?
What is the total disk volume of your data set? (MB)
Is this data set final, unrestricted, and available for release?
What are the reasons to restrict access to the data set?
Has this data set been described and used in a published paper? If so, provide a DOI or upload a digital copy of the manuscript with the data set.
Are the data and documentation posted on a public server? If so, provide the URL.

5.3. ORNL DAAC Curation

After the data provider interaction phase has been closed out, the data set moves into the curation phase. Through the SAuS ingest dashboard, data Quality Assessment (QA) of the data set and documentation assignments are made. During the curation phase, all information collected about the data set is copied and moved to a QA area. The ingest UUID identification is maintained through the curation phase. When the files are moved into the QA area they are piped through a metadata script that extracts file level metadata using open source software such as Geospatial Data Abstraction Library (GDAL), netCDF Operators (NCO), *etc.* This file level metadata is used as a starting point for QA and for building the metadata required for data search, subsetting, visualization, and dissemination interfaces. The file level metadata extracted includes information such as spatial, temporal, file size, file type, variable definition, and associated characteristics of the data files.

QA staff use the information provided through the dashboard, supplemental information collected from the data provider, and the metadata extracted through the script to perform QA checks on the data files. During the QA phase, the integrity of the data files (checksums, projection *etc.*) are verified and the internal and external organizational aspects of the data files (directory structure, file naming conventions, parameter conventions *etc.*) are verified to ensure that the data files are representative of the documentation provided.

QA staff will also evaluate the appropriateness of the file format and make any file format conversions to ensure wider usage of the data files. For example, a binary data file may be converted into a Climate and Forecast (CF) convention compliant netCDF file. The non-proprietary netCDF format and the CF convention ensures that the data are readable many years into the future and allows the data files to be used through a wide variety of data analysis tools. A standards-based file format also allows the files to be readily accessible through web services and other data access, visualization, and subsetting mechanisms, thereby broadening the use of the data files to other disciplines. The data values are never altered during the QA steps and file format conversion process. During the QA process the spatial, temporal, and scientific integrity of the data files are also evaluated. For example, the QA team/person will check if the data files contain the same temporal and spatial extent and resolution as described in the documentation provided by the data provider. In some instances, the ORNL DAAC has received files for a smaller region when the documentation indicates that the data set is global. In addition, the variables described in the documentation are crosschecked with the contents of the data files. For example, in some cases, the data files received may have been scaled but the documentation does not describe the scaling. The QA person identifies such issues to ensure the integrity of the data files and, if necessary, confirms issues with the data provider.

Major issues with the data files are identified during this stage. A detailed QA checklist is provided on the ORNL DAAC website [16]. If there are any unresolved questions or if there are any issues with the data files identified through the QA process, the interaction with the data provider is reopened, and email communications are initiated and tracked to resolve the issues. The speed at which the curation progresses depends on the responsiveness of the data provider and the integrity and completeness of the data files

In addition to the QA, ORNL DAAC staff also prepare metadata for discovery and compile comprehensive documentation that is relevant for future users; we use the 20-years rule [17], a time far enough into the future to be useful for preparing documentation for both sharing and archiving data. Compiling descriptive data set documentation for future users is a time consuming but critical curation process. During the documentation phase, verification is performed to ensure that the documentation matches the files received. During curation, ORNL DAAC staff evaluate if the data set and its contents are clearly described and that the geospatial and temporal information are complete. Other key information about the data files such as the data file parameters, units, research methodology, *etc.* are added to the documentation. If the data set contains data about field stations, information about those field sites (site name, geographic place name, geographic coordinates, elevation, and climate, biosphere, and soil characteristics) is added to the documentation. Calibration information, algorithms, and data quality information are added to the documentation as well. The documentation staff also build a comprehensive reference list that allows users to link the data files to the published research articles that were used to conduct the research and create the data files. Any data use or access policy information is added to the documentation as well.

One of the key benefits of the SAuS system is the ability to centrally manage the documentation and metadata workflow during the curation phase. The documentation is compiled, edited, and approved by several data archive staff. An online metadata editor provided by the SAuS system allows for the documentation and metadata to be shared and edited through a common centralized web based system. The centralized web-based eliminates duplication of document versions residing on individuals' systems and also reduces the need for paper printouts, allowing for editing and approval of the finalized documentation directly within the ingest online system. To keep all of this information synchronized, to facilitate consistency, and to eliminate redundancy the SAuS metadata editor provides views to the metadata XML files, the documentation HTML pages, and the database table view of the data file records. The editor integrates the information across the three views to allow seamless access and eliminates duplication, thereby making the process more efficient. Before SAuS, if the description of a data file has to be updated, for example, the information had to be changed by hand in three places: in the documentation HTML, XML record, and the relational database that powers the

archive web interface. Changes therefore took more time and could possibly have led to inconsistency in content. The centralized automated SAuS system facilitates integration and speeds up the process of documentation creation. Figure 7 shows a screenshot of the metadata editor.

Figure 7. Metadata and documentation editor.

After the quality of the data files is verified and documentation compiled the ORNL DAAC generates a data set citation that includes familiar elements of a citation, including authors, year, title, and digital object identifier (DOI). The DOI for the data set will remain fixed but the location (URL) of the data set may change. The DOI replaces the UUID that was used during internal curation only. An example citation is provided below.

> Thornton, P.E., M.M. Thornton, B.W. Mayer, N. Wilhelmi, Y. Wei, R. Devarakonda, and R.B. Cook. 2014. Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 2. ORNL DAAC, Oak Ridge, Tennessee, USA. Accessed August 25, 2015. Time Period: 1980-01-01 to 1985-12-31. Spatial range: N=35.05, S=32.50, W=-101.80, E=-85.20. http://dx.doi.org/10.3334/ORNLDAAC/1219.

The citation acknowledges the researchers who provided the data products. The SAuS workflow provides the metadata needed to register the data set with the DOI registry using EzID [18]. The metadata used in the workflow also facilitates the creation of a data set landing page. The data set landing page is web location showing access information to the data set made available to a client via resolution of the DOI. The SAuS system also facilitates the creation of the data file and metadata distribution package for user access.

5.4. Ingest Dashboard

The SAuS ingest dashboard is the main operating console for the publication workflow. The ORNL DAAC ingest dashboard has been designed in Drupal [19]. Drupal was chosen for the rapid development requirement for the workflow; in addition, the availability of pre-existing modules for user management made Drupal the preferred development environment. Through Drupal, user management system roles can be easily defined and tightly integrated into the ingest workflow. The SAuS ingest workflow has roles for data set coordinator, ingest coordinator, documentation lead, Spatial QA lead, user services, and chief scientist. These roles can be mapped to individuals or groups

and appropriate workflow permissions can be set within the dashboard. Figure 8 shows a screenshot of the dashboard. The SAuS ingest dashboard also provides various summary information about the status of publication of a data set. Key ingest information can be extracted from the summary information presented through the dashboard. For example, details on ingest time, staff workload, data set priorities and issues, publication timeline, and category of the data set are some information that can be extracted. These details are useful for short-term and long-term planning. A future feature will allow data providers to track the status of their products.

Figure 8. SAuS Ingest dashboard.

5.5. Stage 3 (Publish): Publication and Post-Publication Activities

When the data set is formally released, the ORNL DAAC distributes the metadata to the NASA EOSDIS clearinghouse and other relevant data catalogues. If applicable, the ORNL DAAC also provides tools to explore, access, and extract data. These tools include web services such as the Open Geospatial Consortium (OGC) Web Map Service and Web Coverage Service (WCS), OPeNDAP, and other REST/SOAP-based Web services. The standardization of the workflow through the SAuS system simplifies the integration of the data files into these tools. The ORNL DAAC also advertises the data through email, social media, and the ORNL DAAC website. An automated script prepares an email message to the data authors, congratulating them on publishing the data product and encourages them to add the citation to their curriculum vitae; a DAAC staff member sends this message.

The ORNL DAAC also provides long-term data stewardship for the data set. The archive provides a secure long-term storage of the data files and acts as a buffer between the data users and the data contributors to address any questions about the data set. To provide long-term storage, the ORNL DAAC continuously refreshes its hardware to prevent bit rot and other unintended changes to the data files because of hardware storage issues. The ORNL DAAC creates and tests back-up copies often to prevent the disaster of lost data. ORNL DAAC also maintains at least three copies of the data: the original, an on-site but external backup, and an off-site backup in case of a disaster. In addition, the

ORNL DAAC updates documentation for data sets based on any new information collected. The ORNL DAAC collects and provides data download and citation statistics to gauge the impact of the data sets. Data citations were implemented in 1998 at the ORNL DAAC to provide credit to the data authors, give an estimate of the scientific impact of the ORNL DAAC, and enable readers to access the data used in an article. The ORNL DAAC also added Digital Object Identifiers to its data holdings in 2007 to provide more legitimacy to data citations. The data citations and DOI facilitate the identification of the use of data products in the literature. The ORNL DAAC has integrated data product citations throughout its data workflow and has incorporated data citation metrics to gauge the scientific impact of a data set and to allow data users to understand the various applications of a particular data set.

Figure 9 illustrates an example webpage listing all publications that had used a particular data set from the ORNL DAAC. Example for data set "*NACP Aboveground Biomass and Carbon Baseline Data, V.2 (NBCD 2000), U.S.A., 2000*" doi: 10.3334/ORNLDAAC/1161, http://daac.ornl.gov/cgi-bin/show_pubs.pl?ds=1161 from the ORNL DAAC.

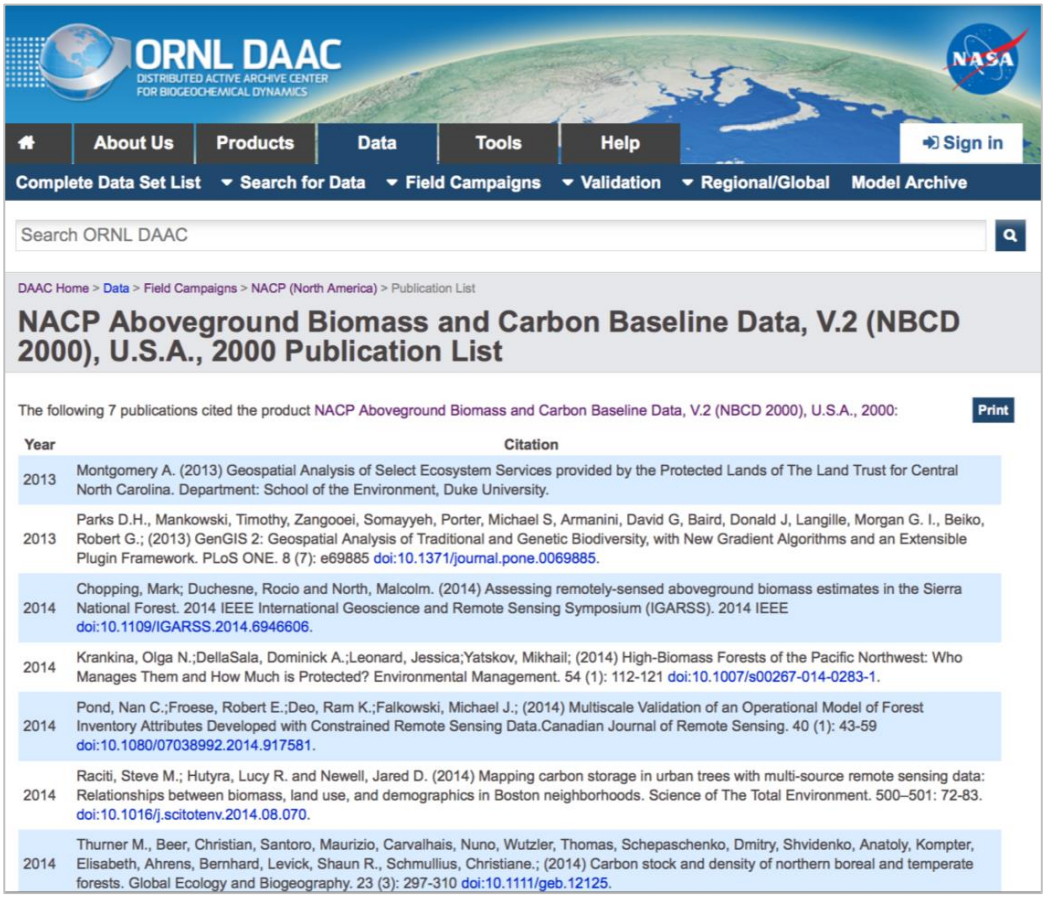

Figure 9. ORNL DAAC website showing publications for a particular data set.

6. Manuscript Publication Process

The semi-automated publication of data sets has become extremely critical for dealing with publication of manuscript-related data sets. Several journals and scientific societies require the archival and unrestricted sharing of the data used in manuscripts prior to publication of the manuscript. This has created a chicken and egg challenge for data centers. The ORNL DAAC, for example, requires that the data provided to the archive have been generated from peer-reviewed literature. With the new journal requirements, data are received prior to manuscript publication. To deal with this challenge, a manuscript publication workflow was added to the SAuS system (Figure 10). In this workflow, a data

set follows the same data provider interaction and curation steps indicated earlier (Figures 5 and 6) but the timing of various steps are triggered based on the timeline of the manuscript publication at the journal. Also, the manuscript workflow incorporates a letter of agreement that details the issues (if any) identified during the data set preliminary QA process. The letter of agreement requests that the data provider responds in a timely fashion to any data related requests and resolves any identified data quality issues. If the terms of the agreement are acceptable to the data provider, a DOI is reserved for use in the manuscript. The reserved DOI will be formally registered during the publication of the data set. A temporary distribution area is also established, in the event that reviewers of the manuscript would like to view the data set. This letter of agreement allows for scheduling the ingest tasks within the context of all tasks at the archive.

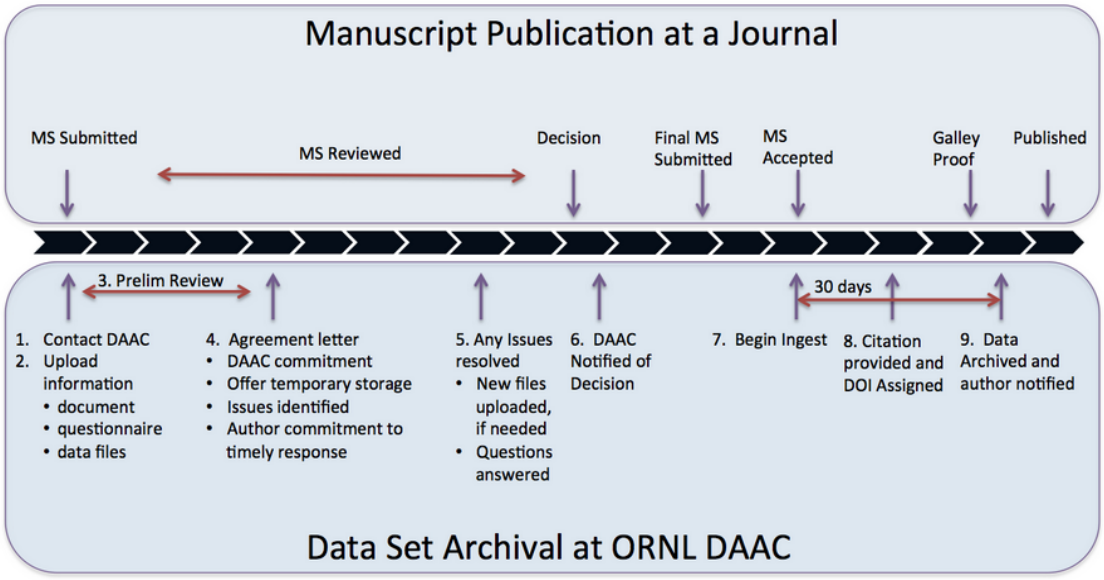

Figure 10. Time-line for archiving data associated with a manuscript. Data products used in a manuscript are submitted to the ORNL DAAC at the same time as the manuscript is submitted to the journal. An agreement is reached between the DAAC and the authors, with the author agreeing to respond to queries in a timely manner, and the ORNL DAAC agreeing to archive the data and provide a citation and persistent DOI close to the time of article publication.

7. Benefits of the SAuS Workflow

For the archive: DAAC ingest staff are required to work with individual investigators on complex and diverse terrestrial ecology data sets. Typically, up to a dozen data sets are in queue for ingest and tracking these data sets are critical for the smooth operations of the archival system. SAuS provides the ability to track a data set from acceptance to publication through a web-based dashboard. SAuS also closely tracks email communications with investigators about a particular data set, so the entire team is able to understand how questions and issues have been addressed, which facilitates further tracking. To improve efficiencies and to reduce redundancy, the SAuS incorporates several automated steps for moving data and metadata files through various stages of the workflow process. Metadata extraction, document generation, and internal communications are facilitated through a publication-tracking dashboard that reduces duplication and improves overall efficiency. SAuS also provides a centralized system of status and metrics, which aids in managing the various aspects of ingest from analysis of ingest time to prioritization of data releases. The centralized file system decreases redundancy and version control and facilitates faster turnaround of data sets. In addition, the workflow system has allowed the ORNL DAAC to update its 20-year-old legacy infrastructure to handle more data sets and improve publication quality.

For the data provider: The SAuS system provides several benefits to data providers. Through the SAuS system, data providers can answer targeted simplified questions about their data set. The targeted set of questions, automated reminders from the SAuS system, and data management guidance documents available on the ORNL DAAC website assist the data provider during the submission. Data providers can also upload data files directly into SAuS. The system provides the ability to track the publication of the data set through the ingest process. SAuS also provides a service to assist authors in meeting the requirements of some journals for archival of data associated with a manuscript [9,10]. This manuscript-related data publication process within SAuS allows the archive and data provider to work collaboratively towards the release of the manuscript and the data set. A description of this process is explained in a later section. After the curation process, an automated email announcement is sent to each of the data authors about the data release. The announcement contains the data product citation, along with a Digital Object Identifier (DOI) for the data set, that can be included in each data author's curriculum vitae. An ability to query the SAuS system to obtain a status report of a particular data set publication is planned for future versions.

8. Lessons Learned in using the SAuS Workflow

The ORNL DAAC started using the SAuS ingest workflow system in October 2013. During the first 24-month test phase of the workflow system, several special cases and refinements were made to the SAuS workflow system to accommodate special cases.

Case 1: Updated Data Files

In this case, a data provider sent updated data files after the initial submission. Updated data files were typically sent after issues with data files or documentation were uncovered during the QA process. In some cases, additional files were added to the submission after the initial submissions were closed out. To handle these cases, the workflow system was updated to preserve the various versions of the data files while allowing the system to reset to an earlier stage in the ingest process. For example, QA had to be repeated after the updated data files were re-submitted to the archive.

Case 2: Data Set Versions

In this case, a newer version of an already published data set was submitted to the archive for publication. SAuS treats the updated version of a previously published data set similar to a new data set. However, documentation, QA, and DOI associations to previously published data set are included during curation and also added to the DOI landing page. An example of a data set with multiple versions is the *Global Fire Emissions Database* data set [20] archived at the ORNL DAAC.

Case 3: Data Provider Question Participation

The data provider questionnaire was designed using input from the user community, but the quality of the answers submitted in some cases has been unsatisfactory. Investigators provided insufficient information for some fields. Also, DAAC staff realized that information in optional fields was too important to be "optional". While the ORNL DAAC received excellent answers for the data provider questionnaire in several cases, updates to the questions themselves are still needed to make them even more useful. Outreach and thorough examples are needed early on in the ingest process to motivate the data provider to submit quality responses.

Case 4: Unresponsive Data Provider

In this case, data providers were contacted after issues with a data set were identified during the QA process. For example, in one data set, the data files submitted by the data provider lacked spatial reference information to correctly project a data set into a Geographical Information System (GIS). The spatial information was critical for correct use of the data. While this was a simple issue that could

be corrected easily, the data provider was unresponsive for our request for the spatial information. In this case, the SAuS dashboard was used to set the data set into a hold status. Some data sets have been in hold status for several months. Further discussion and planning is needed to handle data sets in "hold" status within the SAuS system. One possible option is to send reminders to the data provider through the automated email workflows already available through SAuS. A response can be requested within a few days or weeks based on the nature of the request.

Case 5: Citation Correction after Publication

In this case, a data provider requested that the data citation be updated after the data and DOI were registered and published. Specifically, the data provider changed the list and order of contributing authors after data publication. While the metadata for the data citation could be updated easily, the citation had to be changed in the documentataion and at the DOI registry. The SAuS metadata editor was updated to provide an ability to easily change the citation, including the number and order of the authors in the citation.

Case 6: Approval of the Data Prior to Publication

In this case, the data provider requested the ability to approve the data set's release, or requested that the data be published only after an associated manuscript was published. A task was added to SAuS that allows the data provider to approve their data set just before the DAAC releases it. While this added some additional steps in the ingest workflow, the data author often found items that needed to be corrected, thereby improving the data set. Placing the data provider at the end of the publication step motivated them to be proactive in addressing issues about the data set.

9. Conclusion

The SAuS systems provide a new workflow approach for handling publication of data sets provided by individual investigators. The workflow system maintains the 5-P rule for long-term preservation of the data products, while improving the efficiency and quality of the ingest workflow. The workflow system provided by SAuS can be extended and applied to data publication from any scientific domain. In general, the challenges of data publication including documentation and metadata extraction are similar for all heterogeneous data sets regardless of the scientific domain. Refinements and improvements are still required to make the SAuS workflow more robust. Several data sets have to be tested, and the publication metrics provided by the ingest dashboard have to be analyzed for further modifications and improvements. Although, more modifications are necessary, the SAuS ingest system provides a critical workflow infrastructure for handling data sets efficiently and effectively, especially to address the growing data publication and usage needs.

Acknowledgments: The authors would like to thank Earth Observing System Data and Information System (EOSDIS) for its support, NASA Interagency Agreement No. NNG14HH39I, and all members of the ORNL DAAC data archive for their contributions to the development of the Semi-Automated Ingest System. Oak Ridge National Laboratory is operated by UT-Battelle, LLC for the U.S. Department of Energy, under contract DE-AC05-00OR22725.

Author Contributions: All the authors on this manuscript contributed significantly towards the design and development of the SAuS system described in this manuscript. Suresh Vannan is the lead author for this manuscript. Tammy W. Beaty is the deputy DAAC manager and the project lead for the implementation of the SAuS system. Robert B. Cook is the chief DAAC scientist and provided guidance on the community interaction and in particular the manuscript workflow. Daine M. Wright, Ranjeet Devarakonda, and, Les A. Hook were heavily involved in the technical development and design of the metadata editor and the SAuS documentation elements. Yaxing Wei provided the QA and backend automation design and development. Benjamin F. McMurry provided the scripts for extracting the metadata information into the ORNL DAAC production database. In addition, several ORNL DAAC staff contributed towards the implementation of this system.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kobler, B.; Berbert, J.; Caulk, P.; Hariharan, P. Architecture and design of storage and data management for the NASA Earth observing system Data and Information System (EOSDIS). In Proceedings of the 14th IEEE Symposium on Mass Storage Systems, Monterey, CA, USA, 11–14 September 1995; pp. 65–76.
2. Committee on Geophysical and Environmental Data; Commission on Geosciences, Environment and Resources; Division on Earth and Life Studies; National Research Council. *Review of NASA's Distributed Active Archive Centers*; National Academy Press: Washington, D.C., USA, 1998.
3. Baker, K.S.; Yarmey, L. Data stewardship: Environmental data curation and a web-of-repositories. *Int. J. Digit. Curation* **2009**, *4*, 12–27. [CrossRef]
4. Ball, A. Briefing Paper: The OAIS Reference Model. Available online: http://www.ukoln.ac.uk/projects/grand-challenge/papers/oaisBriefing.pdf (accessed on 20 January 2016).
5. Data Management for Data Providers. Available online: http://daac.ornl.gov/PI/pi_info.shtml (accessed on 3 March 2016).
6. UK Data Archive. Managing and Sharing Data. Available online: http://www.data-archive.ac.uk/media/2894/managingsharing.pdf (accessed on 20 January 2016).
7. Send2NCEI. Available online: https://www.nodc.noaa.gov/s2n/ (accessed on 20 January 2016).
8. National Science Board. Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century. Available online: www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf (accessed on 20 January 2016).
9. Memorandum for the Heads of Executive Departments and Agencies. Available online: https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf (accessed on 20 January 2016).
10. Hanson, B.; Lehnert, K.L.K.; Cutcher-Gershenfeld, J. Committing to publishing data in the earth and space sciences. *Eos* **2015**, *96*. [CrossRef]
11. Nature Editorial. Announcement: Reducing our irreproducibility. *Nature* **2013**, *496*, 398–398.
12. Cook, R.B.; Olson, R.J.; Kanciruk, P.; Hook, L.A. Best practices for preparing ecological and ground-based data sets to share and archive. *Bulletin of ESA* **2001**, *82*, 138–141.
13. Hook, L.A.; Vannan, S.K.S.; Beaty, T.W.; Cook, R.B.; Wilson, B.E. Best Practices for Preparing Environmental Data Sets to Share and Archive. Oak Ridge National Laboratory Distributed Active Archive. Available online: http://daac.ornl.gov/PI/BestPractices-2010.pdf (accessed on 3 March 2016).
14. ORNL DAAC 2016 Archival Priority. Available online: http://daac.ornl.gov/PI/archival_priority.html (accessed on 3 March 2016).
15. Data Provider Questions. Available online: http://daac.ornl.gov/PI/questions.shtml (accessed on 4 March 2016).
16. Data Quality Review Checklist. Available online: https://daac.ornl.gov/PI/qa_checklist.html (accessed on 20 January 2016).
17. Committee on Geophysical Data; Commission on Geosciences, Environment, and Resources; National Research Council. *Solving The Global Change Puzzle: A U.S. Strategy for Managing Data and Information*; National Academy Press: Washington, D.C., USA, 1991.
18. Starr, J.; Willett, P.; Federer, L.; Horning, C.; Bergstrom, M.L. A collaborative framework for data management services: The experience of the University of California. *J. eSci. Librariansh.* **2012**, *1*. [CrossRef]
19. Drupal. Available online: https://www.drupal.org/ (accessed on 20 January 2016).
20. Randerson, J.T.; van der Werf, G.R.; Giglio, L.; Collatz, G.J.; Kasibhatla, P.S. *Global Fire Emissions Database (GFED)*; Version 4; Oak Ridge National Laboratory: Oak Ridge, TN, USA, 2015.