OAK RIDGE
NATIONAL LABORATORY
MANAGED BY UT-BATELLE
FOR THE DEPARTMENT OF ENERGY

ORNL/TM-2013/370

Digital Object Identifiers For OLCF

Version 1.0

September 9, 2013

Terry Jones
Douglas Fuller
Sudharshan Vazhkudai

CONTENTS

Page

ABSTRACT

This document provides an assessment of Digital Object Identifiers (DOIs) and their potential usage within an OLCF context. We present how DOIs can promote more effective scientific discovery for extreme-scale computer facilities charged with supporting wide-ranging research disciplines. We describe the capabilities of DOIs, how DOIs relate to other similar technologies, and how DOIs may be used to facilitate OLCF policies and procedures. Included are the Background and Motivation of traditional DOIs, a technical assessment of DOI strengths and weaknesses, and the potential of DOIs to address expanded requirements within the HPC community.

1. STUDY BACKGROUND

1.1 ORIGINATING MOTIVATION

Recent directives from the Office of Science and Technology Policy (OSTP), the Office of Management and Budget (OMB), and President Obama himself outline a new desire and requirement to provide free access to scientific data arising from taxpayer-funded research [See Appendices C, D, and E]. The provision of this data will require new policies and procedures including a much-improved mechanism for dataset identification and tracking.

A **digital object identifier** (DOI) is a mechanism that can be used to help track and identify the data sets that are produced by researchers globally. To use an analogy, a DOI provides a unique identifier for a piece of data in much the same way as a Bar Code provides a unique identifier for a specific item in the physical world. Like a Bar Code, a DOI may be used by many components (bar code readers, tracking systems, …) and may be employed for many uses once it is assigned. Moreover, the standard helps to integrate systems efficiently. In summary, the DOI is the UPC (Bar Code) for objects of intellectual property on the Internet--it provides a stable, persistent way to uniquely identify "content" (which enables distribution and sharing transactions of all kinds).

The ability to facilitate data-related services gives DOIs the potential for many new and interesting uses. The OLCF could utilize DOIs in their interactions with funding agents by providing *improved accounting* and *visibility* of our user facility production. Also, the OLCF could utilize DOIs in their interactions with our user-community to provide improved mechanisms for citing one's work. Finally, the OLCF can also directly benefit from new 'data strategies' such as data warehousing and other beneficial schemes that result from the improved planning information associated with DOIs and Data Management Planning.

Fig 1. A DOI is analogous to a Bar Code

A DOI is a character string (a "digital identifier") used to uniquely identify an object such as an electronic document. A sample DOI is "10.2224/2003-1-29-CENDI-DOI". Metadata about the

object is stored in association with the DOI name and this metadata may include a location, such as a URL, where the object can be found. The DOI for a document is permanent, whereas its location and other metadata may change. Referring to an online document by its DOI provides more stable linking than simply referring to it by its URL, because if its URL changes, the publisher need only update the metadata for the DOI to link to the new URL.

The DOI website is http://www.doi.org/

The datasets (datastreams, data files, etc.) produced at scientific user facilities like the OLCF support a wide range of technical reports and published literature. They are also recognized as valuable information entities in their own right. The directives from the current administration speak to the desire to make these datasets available for citation, discovery, retrieval, and reuse. This challenging task may benefit from the *assignment and registration* of a DOI for every dataset; each such DOI is available via a free service for the OLCF user community. The free *assignment and registration* service is provided by OSTI to enhance DOE's management of this important resource.

When a DOI that permanently links to the dataset's location is registered, it is disseminated to databases and to commercial search engines such as Google via the RA (OSTI/DataCite).

1.1.1 OLCF Benefit 1 – Helps with 'Research Tracking'

DOIs will enable the OLCF to more accurately assess the production of projects.

1.1.2 OLCF Benefit 2 – Helps with 'Reporting To Sponsors'

DOIs will also enable the OLCF to provide sponsors with improved information about their investments.

1.1.3 OLCF Benefit 3 – Helps resolve data disposition questions

DOIs will also enable the OLCF to know what to do at the end of the project, and will decrease the likelihood of unwanted purge loses.

1.1.4 OLCF Benefit 4 – Provides many benefits To Our User Community

The traditional benefits of DOIs are described in section 1.4.

1.2 A SAMPLE DOI

The following example is taken from the Carbon Dioxide Information Analysis Center (CDIAC) at ORNL which is a NASA funded project. As NASA has long had the mindset of making their various sensor data permanently available to researchers, they are ahead of DOE in regards to DOI deployment.

DOI taken from CDIAC Database at ORNL	
Full reference	Boden, T.A., G. Marland, and R.J. Andres. 2012. Global, Regional, and National Fossil-Fuel CO_2 Emissions. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tenn., U.S.A. DOI 10.3334/CDIAC/00001_V2012
DOI	10.3334/CDIAC/00001_V2012
Resolves To:	http://cdiac.ornl.gov/trends/emis/overview_2009.html

Table 1: A Sample DOI

1.3 COSTS INCURRED BY OLCF TO DEPLOY DOIs

This section describes what the OLCF is signing up for if they elect to provide a DOI service.

1.3.1 DOI Budget Impacts to OLCF

Primary funding for DOI usage within the OLCF is an issue as the data will need to be maintained past the life of the creating project (the "data mortgage" issue). One potential way to fund the data mortgage is to add funding for such expenses with the original data-producing projects (which will soon be required to provide Data Management Plans anyway).

Incorporating a DOI-based service within the OLCF incurs the following budget considerations. [Note that the incorporation of Data Management Planning will make it much easier to right size resource capacities, and will decrease the likelihood that data intended to be preserved by the user is deleted by the center.]

- The provision of a web host for the front end (currently doi1.ccs.ornl.gov)
- The provision of sufficient HPSS storage to archive approved/planned data from OLCF's user community.
- Additional HPSS storage if we decide to expand the facility beyond data generated at OLCF.
- The network bandwidth to accommodate ingest, distribution, and catalog queries.
- The necessary manpower to manage these resources. While none of the resources are new (web host, HPSS, networking), each will need to be increased in capability or capacity.

1.3.2 DOIs Require The Citable Data To Be Available Permanently

DOIs are global and expect persistent data, available at the hyperlink provided as part of the DOI data. In brokering a DOI request, OLCF obligates itself to permanently store, track, and make available the referenced dataset online without a clear time limitation. This may be implemented by enabling external access to archive storage (i.e. HPSS).

1.3.3 DOIs Require Associated Metadata To Enable Effective Searching

While many of the fields within the DOI metadata are optional, effective use of these optional fields is critical for search and discovery. Researchers applying for DOIs will need to be educated on the merits and requirements of this process. OLCF may wish to define local policies for required metadata fields, description tags, and keywords.

1.3.4 DOIs Require That The Data Has Data-Integrity

As part of storing and making available data referenced by DOIs for an indefinite period, data integrity becomes increasingly critical. Uncorrected data errors, even those that are detected, reduce or eliminate the science impact of the dataset. Additionally, providing known-bad, unknown-bad, or otherwise faulty data is damaging to the reputation of OLCF.

1.3.5 DOIs Imply Data Management Planning

The above requirements imply a need for increased long-term storage capacity, integrity, and availability. This will require significant direct expense as well as regular upgrades and other overhead costs such as system administration. These data management activities will require effective planning to ensure sufficient resources and manpower are available to support the effort.

1.4 BENEFITS OF DOIs

1.4.1 Promotes Science

Announcing and registering datasets with DOIs enables researchers, especially future researchers, to more easily discover the data, access it, and reuse it for verification of the original experiment or to produce new results with the latest methods.

Furthermore, datasets that have been announced and registered become searchable in OSTI's databases, including Energy Citations Database, Information Bridge, and the DOE Data Explorer. Users of these databases are linked to the dataset at the data center or facility where it resides; this increases the opportunity for discovery of additional data, specialized interfaces, toolkits for data analysis, etc.

1.4.2 Promotes The Data

Because of the responsibilities a submitter must meet in order to have DOIs assigned for datasets, users seeing those DOIs know the information has a level of integrity and commitment backing it that becomes part of its provenance.

Because OSTI is the operating agent for Science.gov and World Wide Science.org, datasets become searchable there also; and, due to the agreements OSTI has in place with commercial search engines such as Google, your data becomes visible to their users as well.

1.4.3 Quantifiable Measure of Enabled Research

DOIs facilitate accurate linkage between a document or published article and the specific datasets underlying it.

1.4.4 Unequivocal Identification

DOI's make data easy to cite in a standardized way [DOIs have become recognizable as pointers to important information around the globe], encouraging authors to include this step in their writing/publishing activities. Unique and reliable.

1.4.5 Permanent Identification

Because a DOI link is a persistent link, unlike a URL, publishers and others who use DOIs create reliable, persistent links in citations and database records. No broken links.

DOI's can always be "resolved" at DataCite. Anyone seeing the DOI in a print publication, for example, can access the DataCite homepage, type the number into the resolver tool, and find out what information that DOI is identifying, where the information is, and then go there with a click of the link. And, of course, online publications will list DOIs as live links when authors reference their own datasets or those of other scientists.

1.4.6 Associates Data and Metadata

Uses the indecs Content Model. Adds value to electronic publications: Readers have come to expect online material to contain outbound links to cited sources. At the same time, DOI linking will augment the accessibility of content through inbound links

1.4.7 Widespread Support

Designed to be human readable, based on open architecture. A single DOI serves as a linking agreement with all participating publishers. Avoid having to sign numerous bilateral linking agreements with publishers.

1.4.8 Extensible Support

Designed to operate flexibly.

1.5 DOI ORIGINS

DOIs are the result of the International DOI Foundation (IDF), or http://www.doi.org. This is an open member organization launched in 1998. Among the membership of the IDF are publishing companies, technology companies, and intermediaries. The DOI was modeled on W3C, and on the Bar Code development.

.

1.6 DOI DRAWBACKS

- Lack of API support beyond POSIX
- Machine-dependent implementation
- Limited extensibility
- Reduces number of clients, but not number of operations.

1.7 MAJOR DOI INSTITUTIONS:

1.7.1 International DOI Foundation (IDF), a non-profit founded in 1998.

1.7.2 Corporation for National Research Initiatives (CNRI)

2. TECHNICAL ASSESSMENT OF DOI DEPLOYMENT FOR HPC

2.1 LOGISTICS OF DEPLOYING DOIS

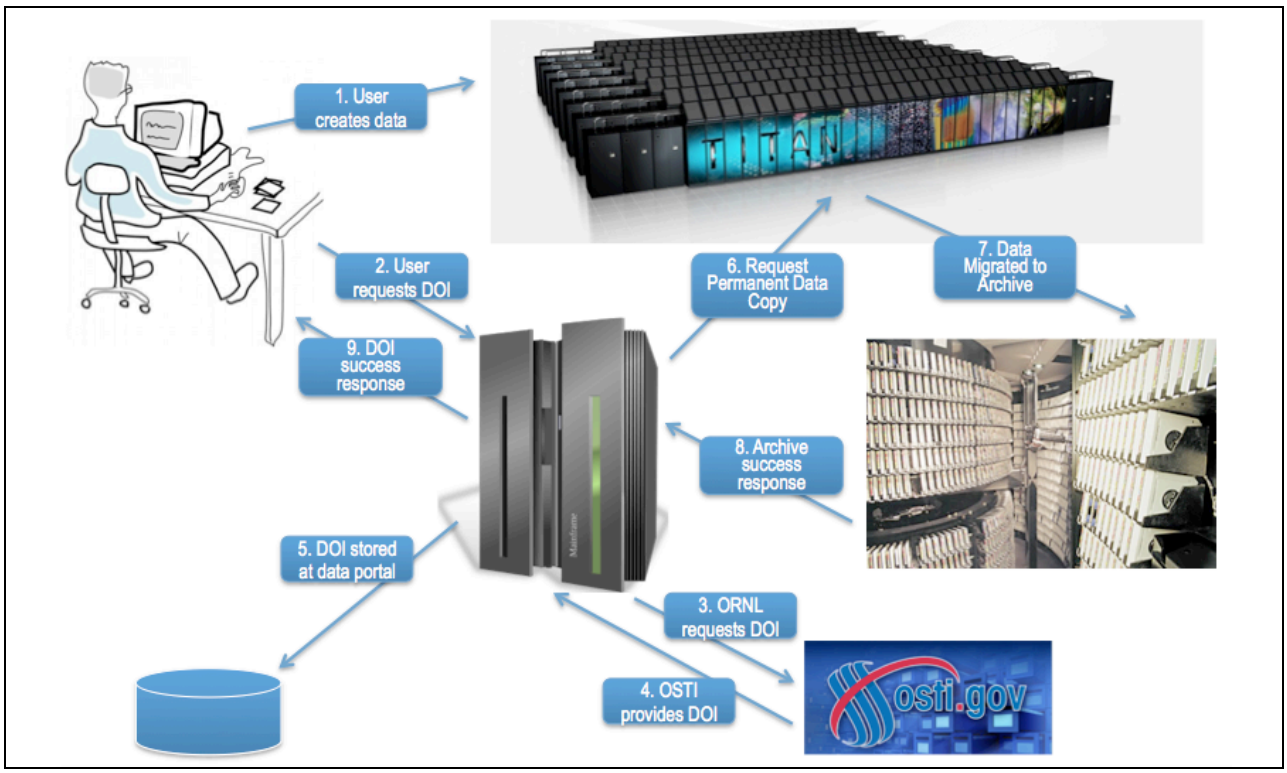


Figure 2. Workflow for DOI Creation

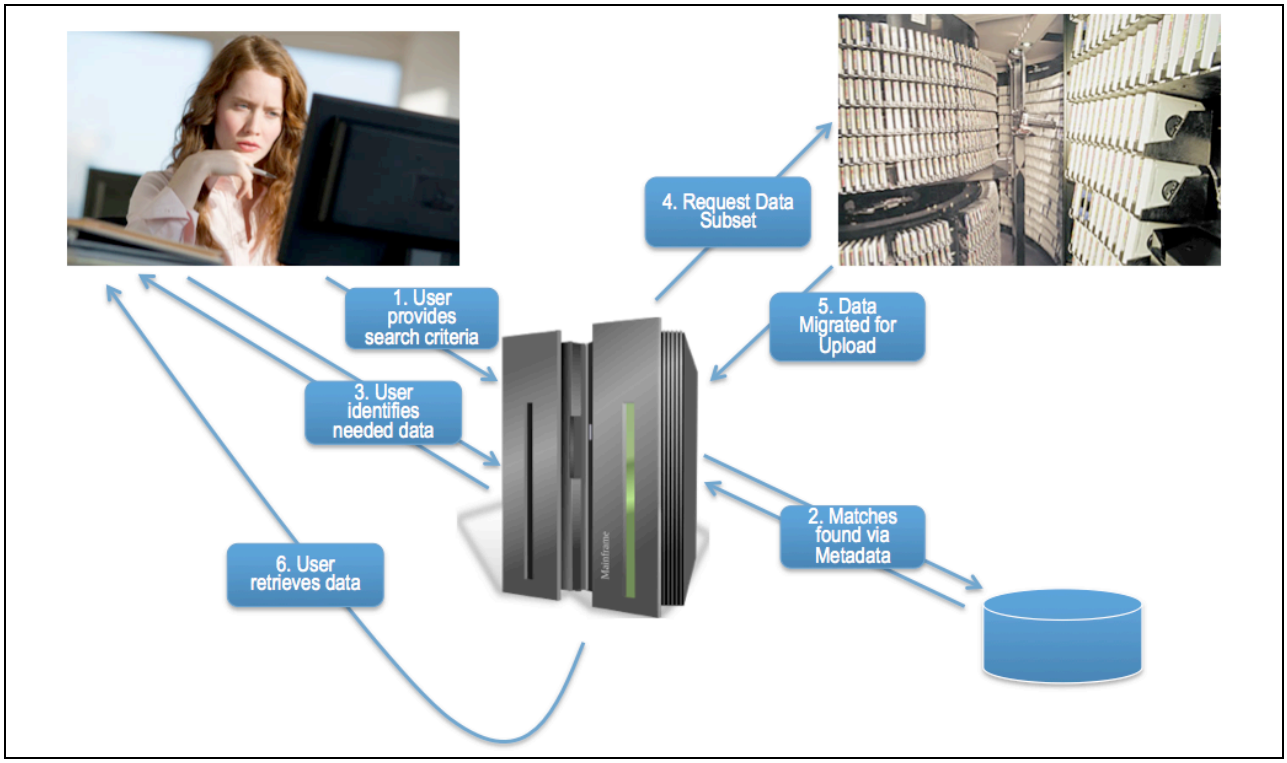


Figure 3. Workflow for DOI Data Retrieval

Figure 4. Frontend design

2.2 A COMPARISON OF DOI REGISTRATION AGENCIES (RA)

	Coverage	Pros & Cons
OSTI.gov (through DataCite) Mark Martin martinm@osti.gov 865-576-2097 Jannean Elliott ElliottJ@osti.gov 865-576-6784	• The collection must consist primarily of non-text information, such as numeric files, figures or data plots, images, multimedia, etc. Some collections include text but only as a part of a specialized mix. The content types are defined in Help. • The data must be the result of research and be maintained for reference purposes, analysis and reuse, or in support of specific projects. Sample data and mere calibration data were excluded, as were operating statistics and normal log data for DOE's many research instruments. Specialized tools and codes may be part of the data collection, but collections that are primarily toolkits and software were excluded. The exception is computer models and simulations. The line between the tool and the tool's results can be blurry in some of these cases. • The collection must be funded either wholly or in part by DOE. There must be evidence of DOE financial support that either caused the data collection to be generated or now helps in the maintenance of the data collection. • The collection may be small but should consist of more than just two or three items. Multiple items must logically fall under the collection's "title."	– Doesn't provide ARKs + office in Oak Ridge + free
EZID (through DataCite) Joan Starr Joan.Starr@ucop.edu 510-987-0469 John Kunze John.Kunze@ucop.edu 510-987-9231 (http://n2t.net/ezid)	• You want to **cite the dataset right now** even though you haven't yet found a permanent "home" for the data. (It's still on your desktop.) • Or, maybe you don't even have any data yet. You can get a preservation-ready identifier, such as an ARK or a DOI, right now with EZID that you can use in your paper. • Store your data anywhere. Use EZID to update the location details. • Move your data easily when members change institutions. • To meet funding requirements like a data management plan (organizing and naming files)	+ Provides ARKs. ARKs do not come with a metadata standard; you can use your own. + Provides some intermediary services that come into play before the submission to DataCite. These allow you to choose your own DOI suffix. – $2500/year from now on since DOIs are forever (although we could use the Mercury Consortium)
ORCID (orcid.org)	Just announced DOI registration authority for APS. EZID and OSTI identifiers are assigned to data or resources. ORCID iDs are created for individuals. An EZID or OSTI number stays with a single resource as an identifier, while an ORCID ID can be added to any resource as a way to identify the creator. But users can add works they have created to their ORCID records and our system accepts OSTIs and DOIs for identifying these works. ORCID also differs from other identifier for researchers by being interoperable with numerous systems and seeking less to provide a researcher profile but instead to have an identifier that can stay with users throughout their researcher career.	+ ORCID is partnering with publishers so that an ORCID iD can be included with author names in academic articles and will soon be working with universities and granting organizations too. – ORCID is only 8 months old, so we're just starting to have ORCID iDs accepted into more and more systems

Table 2. Comparison of Registration Authorities

2.3 ABOUT DATACITE

- Focused on improving the scholarly infrastructure around datasets. There will be a set of activities around establishing and sharing best-practices, identifying and solving some of the unique issues that arise with datasets.
- Focused on working with data centres and organisations that hold data. The details of their business models, workflows, and other requirements do not appear to be identical to those of publishers producing traditional journals.
- Has a business model that meets the needs of non-commercial and sometimes smaller organisations; larger national-scale organisations (e.g., TIB, BL) carry the basic infrastructure costs and will reclaim where appropriate within their domain.
- Participates in A DOE researcher, organization, or grantee determines that important datasets exist which in EPIC.

2.4 HOW OLCF WOULD INCORPORATE DOIS

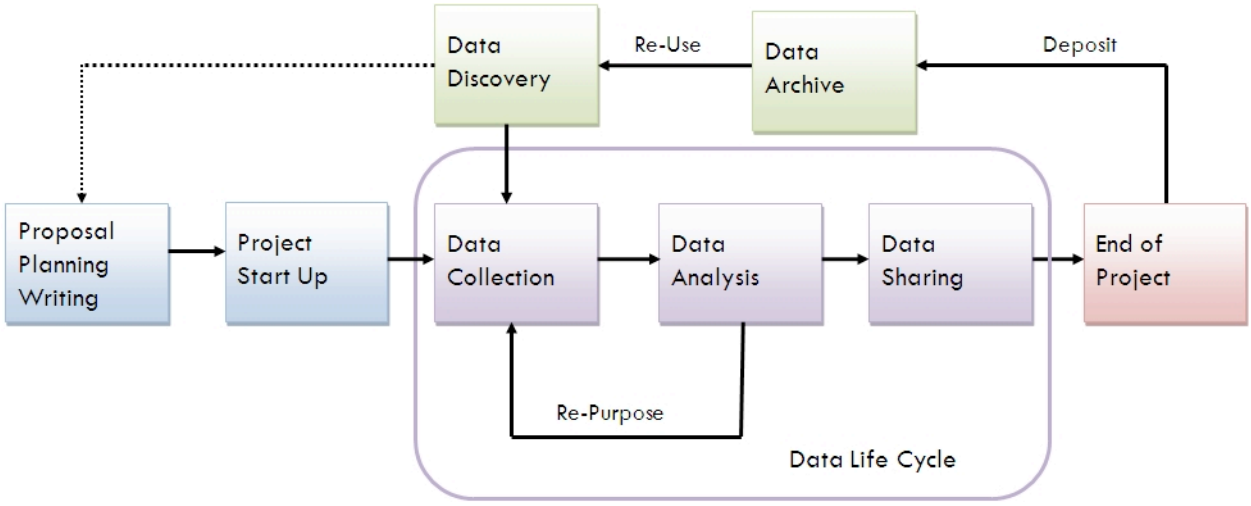


Figure 5. Data Life Cycle

2.4.1 Logistics / Process of Generating A DOI.

1. A DOE researcher, organization, or grantee determines that important datasets exist which need to be announced in DOE's scientific and technical databases and assigned DOIs. DOE Order 241.1B instructs that bibliographic information for these datasets be submitted to OSTI. First time submitters may contact OSTI at 865-576-6784 for help in deciding what submittal method will be used, metadata requirements, etc.
2. Submittal is handled through manual input into the Announcement Notice (AN) 241.6 or via an automated 241.6 web service/API. The AN 241.6 is available without login or through an E-Link account for DOE organizations. It is appropriate if the anticipated volume of datasets to be registered is low. Higher volumes are more easily handled through the automated API, but some upfront programming is required of the submitting site.
3. The submitter decides at what level DOIs will be assigned to the data. Some datasets are similar to collections, in that they have multiple data files to which the "landing

page" leads; others may be small and consist only of an Excel spreadsheet. Defining the boundaries of the datasets that will be announced and registered is an important step requiring subject expertise and knowledge of how particular audiences normally look for the data in question. For that reason, the definition of what constitutes a dataset that will receive a DOI is the responsibility of the people who know the data best, i.e. the people at the submitting organization.

4. When announcing and registering a dataset, the submitter agrees to the following requirements:
    - Ensures that the dataset is located in a data center or online repository where it will be managed in such a way to provide persistent access and maintenance of all URLs associated with the DOI.
    - Provides, at a minimum, the mandatory metadata and ensures appropriate authority to make available the metadata and the dataset being identified. The dataset must be open and accessible to the public.
    - Ensures that the URL assigned to the DOI links to a "landing page" (typically an HTML page) that provides users with the necessary context for using the data.
    - Coordinates with OSTI to create and maintain a persistent "tombstone page" when data registered with a DOI must become unavailable.

2.4.2 Details of Interfacing With OSTI

Manual Submission

- If the AN 241.6 will be manually completed and submitted, the steps for doing so are below. Go to #6 for using the automated 241.6 web service/API.
    a. Log in to E-Link with an existing account or request a new account at https://www.osti.gov/elink/register.jsp (AN 241.6 on E-Link) or simply begin entering metadata at https://www.osti.gov/elink/241-6.do?ostiid=0&action=load (non-login version).
    b. Ensure all required fields are completed with the appropriate information. Required fields are: Dataset Type, Dataset Title, Creator/PI name, Dataset Product Number, DOE Contract Number, Originating Research Organization, Publication/Issue Date, and the URL for the HTML landing page. Some minimal contact information is also required for administrative purposes. More detailed information is provided at https://www.osti.gov/elink/F2416instruct.jsp?printerfriendly=true
    c. Provide information for as many of the optional metadata fields as possible. A description of the dataset is optional but is highly encouraged because of its extreme importance for enhancing searchability.
    d. Use the buttons at the bottom of the AN 241.6 to submit the metadata to OSTI. Note that the dataset itself is not physically transmitted to OSTI, only the metadata that identifies and describes it. OSTI does not upload or host the actual dataset.

OSTI's processing of the AN 241.6 metadata now begins. Two unique numbers are assigned: (1) the OSTI ID that identifies the record in any of OSTI's databases and (2) the unique DOI that will identify the dataset and its location to the world. Both numbers are immediately available to the submitter through a printable confirmation page and an automatically-generated email. The

format of the DOI will be a numerical string beginning with 10. and continuing with a / (slash mark) and a number known as the "prefix," then another / (slash mark) and the OSTI ID. Example: 10.5439/1023895 or http://dx.doi.org/10.5439/1023895.

Automated Submission

- If the submitter wishes to utilize OSTI's 241.6 Announcement Web Service/API, the steps for doing so are:
  a. Obtain account information from OSTI by calling 865-576-6784.
  b. Have an IT developer at your site write the software routine that will connect to OSTI's 241.6 web service/API. The developer will also write the code that will pull metadata from your database or location where it resides, will format it into correct XML file format, and will then transmit the file to OSTI's server as a "Post" operation. For more detailed information about the programming steps, including sample code and sample XML records, see the 241.6 STI Announcement Web Service Manual.
  c. Required and optional metadata fields are the same for all entry methods. See 5.c and 5.d in this document or find detailed information about the metadata in the manual referenced above.
  d. Coordinate with OSTI to request OSTI's test systems. Once the routines are working smoothly, OSTI sets up a production account for the submitter, and the metadata flows directly to E-Link, OSTI's processing system. The schedule is determined by the submitter. Metadata can be submitted once daily, several times a day, once a week, etc.

OSTI's processing of the metadata begins as soon as a transmission is received. Two unique numbers are assigned: (1) the OSTI ID that identifies the record in any of OSTI's databases and (2) the unique DOI that will identify the dataset and its location to the world. Both numbers are immediately available via the XML response that OSTI's server returns to the submitting server and through an automatically-generated email. The format of the DOI will be a numerical string beginning with 10. and continuing with a / (slash mark) and a number known as the "prefix," then another / (slash mark) and the OSTI ID. Example: 10.5439/1023895 or http://dx.doi.org/10.5439/1023895.

2.4.3 Metadata Requirements For A DOI

- Dataset Type,
- Dataset Title,
- Creator/PI name,
- Dataset Product Number,
- DOE Contract Number,
- Originating Research Organization,
- Publication/Issue Date,
- URL for the HTML landing page
- Some minimal contact information is also required for administrative purposes

3. RESULTS OF DOI DISCUSSIONS WITH OLCF USER COMMUNITY

3.1 SAMPLE TABLE OF USER-COMMUNITY DISCUSSIONS

	Saved Data • input decks? • output decks? • other?	Granularity of DOIs	Community Receptiveness to DOIs	Number of Files & Number of Bytes
Stellar Astrophysics (Bronson Messer)	No on input deck (no community code) Output is possible: very grid based (turbulence & strong gravity); optical and other spectra are possible (but not as clear)	Mixed bag. There are simulations that are big and unique enough to merit a DOI; in others, it would be for a suite of simulations tied together with some common physics component.	Likely would not initially see the benefit. There's not much public release of data. Hard to gauge, but likely lower than cosumology.	20,000 files for the suite of files associated with a DOI. 1.2 to 1.5 TB. **future:** more dimensions (500x bigger than now to 3D) to real transport (total 1000x bigger than now)
Cosmology (Bronson Messer) mpe-garching,mpg.de/millineum	* input decks (lots of community codes); sometimes there is special sauce (initialization) * very grid based, plus synthetic observations (ripe)	Simulation campaigns are big enough that they name individual runs. A "simulation" would be suitable for its own DOI.	Not bad. Reasonable chance that people would use them.	One database. Total 20 TB for a project (single simulation, ran over many well times) performed in about 4 to 8 weeks. **future**: not looking the number of particles, but will be increasing the fidelity of the transport (perhaps 100x bigger)
NOAA & Turbulence (Duane Rosenberg)	Direct Numerical Solution -> not many fudge factors (small input decks); output is shareable and is of general interest.	A "simulation" would be suitable for its own DOI.	Already in use (Johns Hopkins)	solutions are co-located on 3d meshes with perhaps 7 scalars per mesh point. Large is 8000^3 or 4 TB **future:** Scale with memory.
Fusion / Particle in Cell (David Green)	For each scenario, (that starting point) you don't start from first principles and couple everything together. The input deck is of use to others.	One DOI for a project page, and then you get the description and things that go with it. Wouldn't want it to be too small.	Simulation people (especially younger ones) will like it a lot; but the institutions may require some effort in how they handle their policies.	150 GB for the particle data; 15 GB for the mesh data; 1000 dumps (one run) is 160 TB

Table 3. Sample User Comments on Data Practices

3.2 ASTROPHYSICS DETAIL – BRONSON MESSER

3.2.1 Fields which have significant data and/or workflows

(a) Climate which has both turbulence and sensor data;
(b) Molecular Dynamics which has turbulence;
(c) Fusion.

3.2.2 Other Fields not as suitable for Data Warehousing

(i) Chemistry and Materials Science;
(ii) Material Dynamics;
(iii) Nuclear Physics;
(iv) most engineering-based disciplines.

3.2.3 Bronson's recommended contacts

(a) Climate -- Anantharaj Valentine; also NOAA & Turbulence -- Duane Rosenberg;
(b) Turbulence -- Ramanan Sankaran;
(c) Fusion Particle-in-cell -- David Green;

3.2.4 The Virtual Astronomy success story

Everyone is able to develop input decks based on star data because of the use of a consistent schema for that data.

3.2.5 Any Known User's of DOIs?

Bronson speculated that Astro already uses DOIs for data. He believes one example would be the Palomar Plates.

3.2.6 Other Recommendations

Carrot & Stick approach to data: (a) carrot -- avoid the purge; (b) stick -- funding sponsor restrictions.

3.3 GROUNDWATER SIMULATION DETAIL – BOBBY PHILIP

3.3.1 Workflow Overview

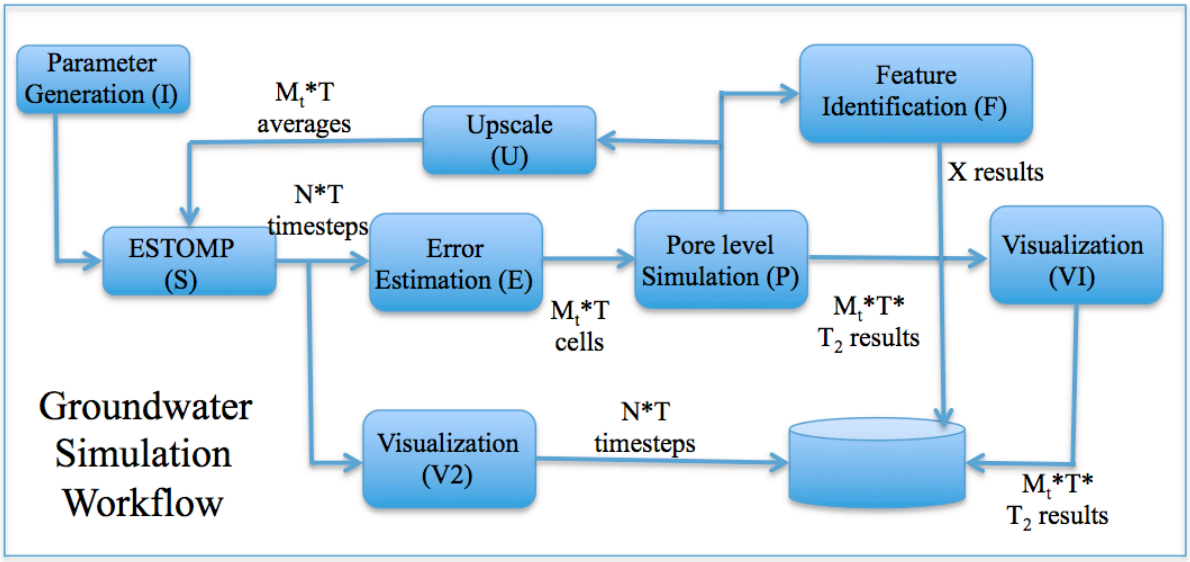


Figure 6. Groundwater Simulation Workflow

The groundwater workflow, shown in Figure 1, is an example of an adaptive multi-scale physics problem with in situ data analysis and visualization included. The workflow starts with the generation of a set of N input parameters representing N different geospatial distributions of materials, where N is typically between 30 and 100 but can go up to 1000. Each input parameter is used to launch an ESTOMP simulation (S), which is a parallel simulation running on over 100k cores. Each ESTOMP run performs a continuum hydrodynamics simulation comprised of 100M – 500M cells and generates T (hundreds to thousands) timesteps of mesh data. Each mesh is evaluated to determine if the error in any of the cells is above a threshold. If a cell is above the error threshold, and it is expected 1-5% of the cells will be, S is paused and a set of M fine grained pore level simulations (P), which is either a smooth particle hydrodynamics code or a computational fluid dynamics code, are called to refine the simulation in those areas. The number of cells that need to be refined changes on each timestep and is not known in advance. Each P is itself a parallel simulation running on ~4k cores, for T2 timesteps. Once a P converges to an answer, the results are upscaled and returned to the ESTOMP code, which then moves to the next timestep. Concurrent with the simulation execution, analysis routines create visualization of the data and save those to disk. An analysis step (F) can also be performed to determine if any output of P should be permanently saved. The N simulations generate a collection of results which are analyzed off-line to determine the sensitivity of the problem to different material distributions and the overall results distribution. S and P will never be running simultaneously, so they can reuse the same cores (if the goal is to increase computational efficiency) or not (if the goal is to minimize data transfer). If one of the Ps encounters a fault, it can be restarted from its initial conditions. However, if S encounters a fault, it should resume from the previous checkpoint file and all the Ps executed since that last checkpoint should be rerun.

3.3.2 Groundwater Workflow Extended For DOIs

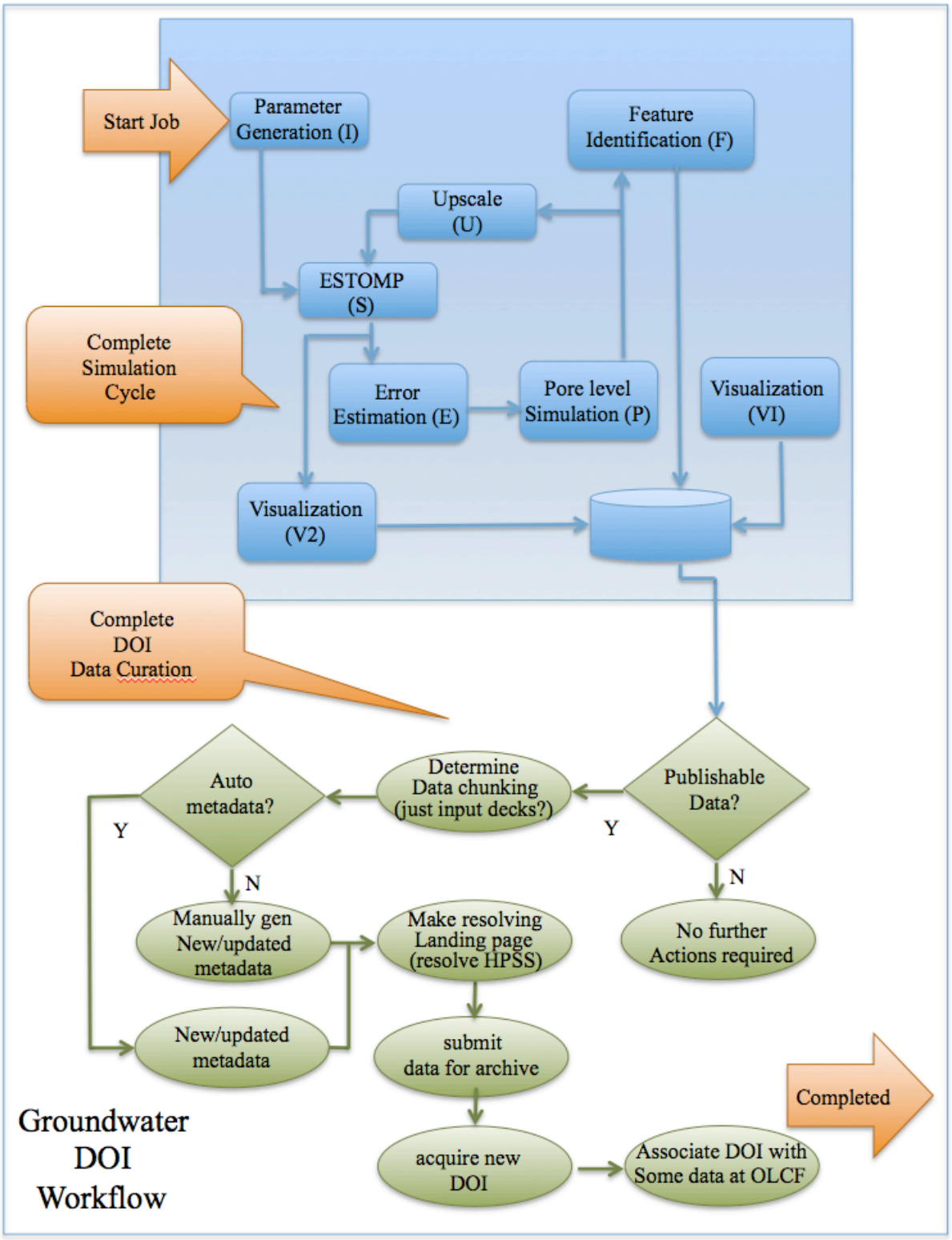


Figure 7. The steps for acquiring a DOI for Groundwater Simulation

3.4 CLIMATE DETAIL – VAL ANANTHARAJ

3.4.1 Climate Workflow Overview

For some time now, the Climate community has been doing DMP plans for NSF and NASA. Their primary work is in support of Earth Systems Grid (ESG). DOE is a primary Climate Research sponsor, but the field in general is funded by an international group of agencies. But the nature of NSF and NASA has been predominately geared toward sensor measurements (which are retained forever), while DOE has funded its efforts through activities like SciDAC and other ASCR-funded projects which have a sunset (and thus no continuity).

The predominant code within the Climate Research community is CESM (Community Earth System Model). The PI comes up with "experiments" to run against CESM. These experiments are not short term like weather – it's multi-decade simulations.

The workflow does not contain feedback loops as in groundwater research. Climate Scientists do not interact or iterate once they have set up the model. There is no user interaction. The only workflow adjustment may be the inclusion of "validation scripts". After one completes set up for the grid, the entire experiment proceeds with that specific model. A typical experiment is "frozen" for the entire length of the experiment. This is to support sensitivity experiments.

Visualization is always there as a post-processing step (for publication).

As of 2012, this workflow is now supported by the various resources available at the Oak Ridge Leadership Computing Facility (OLCF) [Anantharaj2012]. The available systems include the 2.3 PF Jaguar Cray XT5 supercomputer, the 10 PB Spider center-wide parallel file system, the Lens/EVEREST analysis and visualization system, the HPSS archival storage system, the Earth System Grid (ESG), and the ORNL Climate Data Server (CDS). The workflow enabled on these systems, and developed as part of the Ultra-High Resolution Climate Modeling Project, allows users of OLCF resources to efficiently share both simulated data, often multi-terabyte in size, as well as the synthesized products derived from these datasets.

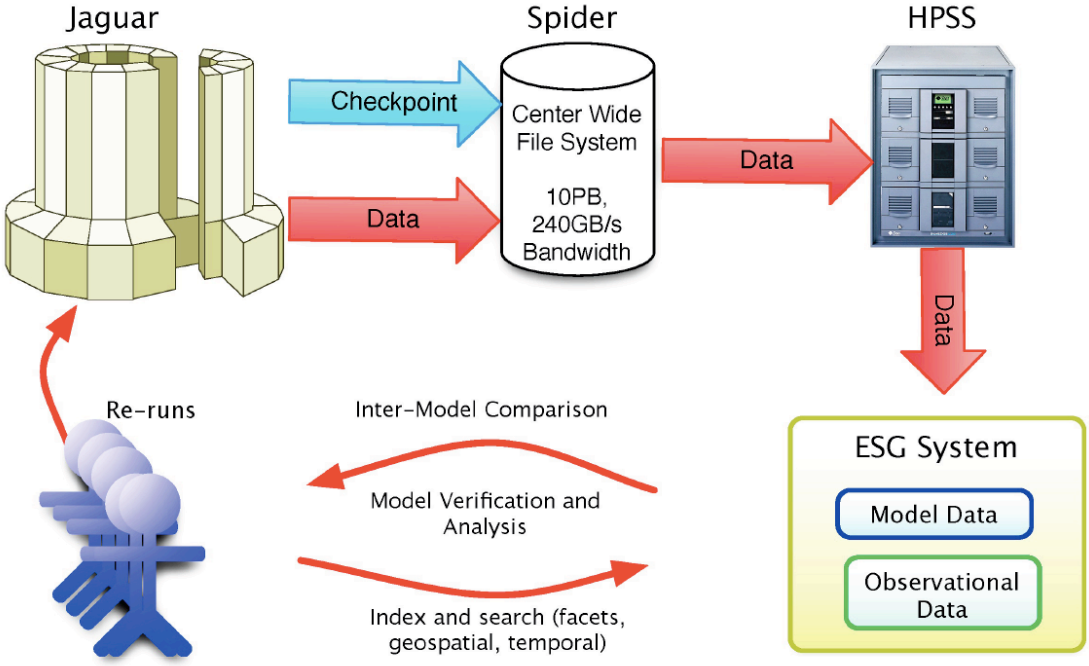


Figure 8. A sample Climate Workflow

3.4.2 Should data be preserved?

In general, perhaps only 5% to 10% of climate data should be retained for posterity.

With current computing power, the necessary resolution to get "interesting" results requires too much computational power for direct-numerical-simulation methods. Hence, the state of the art is in developing accurate heuristics to fit the desired resolution on the available machine. These heuristics include many adjustable parameters and/or adjustable regime points. Each output dataset therefore requires extensive metadata to describe in enough detail the strategies employed to generate the data. Models are vetted through a community vetting process. Results are compared against "mean state" of the climate.

These heuristic techniques are constantly improving. If a result is too old, it may be worthwhile to recreate the data with the superior techniques of today. A typical transition point might be "was this calculated over 10 years ago?"

The progressive nature of these heuristics mean that the input decks (which would contain the various adjustable parameters and regime transitions for a specific model) are not of long-term interest.

3.4.3 Is Data Applicability Wide Or Narrow?

The model folks and the folks developing physics in CESM are for the most part separate. CESM is a framework that includes many models (land, ocean, atmosphere).

3.5 NOAA & TURBULENCE DETAIL – DUANE ROSENBERG

3.5.1 NOAA & Turbulence Workflow Overview

Unlike some other domains (e.g. groundwater research), the turbulence research community work often with direct-numerical-solution (DNS) problems which leads to a much simpler workflow in terms of simulation production. Significant effort might be spent on "setting up the problem" (by running lower resolution simulations or other calculations), in order to choose the equation parameters (diffusivities, rotation rates, stratification, resolution), but then the solution follows during production in a specific and unambiguous way from the problem setup. This set up step is not a 'tuning' step as in the case of cllimate modeling, or even large eddy simulation (LES); it simply enables specification of physics terms exactly in the underlying partial differential equations being solved. DNS allows researchers to capture all spatial and temporal time scales in a flow with high accuracy. The solutions are co-located on 3d meshes with perhaps 7 scalars per mesh point. Universities typically solve problems around 512^3 to perhaps 1024^3 (that's stretching it). Big machines like Titan solve problems of perhaps 2048^3 to maybe 5000^3, and can handle perhaps up to 6000^3.

Large runs permit investigations of large Reynold's number flows. [Low Reynold's number indicates that a flow is dominated by viscous forces and is therefore laminar. High Reynold's number indicates flow is dominated by inertial forces which tend to produce chaotic eddies, vortices, and other flow instabilities that are indicative of classical turbulence.]

3.5.2 Should Data Be Preserved?

For small problems (those under 512^3), it's probably easier to recreate the data, unless the data was a parameter study or involved, say, exceptionally high output frequency. For larger problems,

the general usefulness of the data and the tremendous effort in generating it dictate reuse whenever possible. Therefore, there is recognized benefit in retaining all "difficult to come by" result datasets.

3.5.3 Is Data Applicability Wide or Narrow?

The resulting NOAA/Turbulence output datasets contain a wealth of information. Publications typically present info on various statistical analyses of the output datasets, or some kind of reduction technique associated with a particular scalar or vector field. Many of the classical 'statistical' analyses (e.g., spectra, spectral transfer, conserved quantities) are most easily computed during production, and can be provided at a much finer time resolutions than the volumetric data can be. But, strictly speaking, anything that _can_ be derived from the equations is contained within the volumetric data (assuming sufficient time snapshots); the upshot is that one dataset is a goldmine for various researchers to try their analyses. As such, a professor at Johns Hopkins has started a data repository of these datasets (funded by NCF) – see Appendix C.

The current datasets at the Johns Hopkins website are small, and it would be very useful to have some kind of subsetting features (or nearline analysis tools if work is to be done in situ).

3.5.4 Projected Future Data Requirements

The exascale machines are expected to solve problems of size 8000^3. Further data to be obtained from the DMP form from Duane.

3.6 TURBULENCE – RAMANAN SANKARAN

3.6.1 Turbulence Workflow Overview

When those in the turbulence community generate data from the large runs, they analyze the results and write papers. These are usually accounted for since they are reported to OLCF through quarterly reports and such.

There are more papers that come from the outside community that publish their own analysis of the data. Some of these groups want to start from the full multi-TB dataset and run their own analysis. There are also papers that just use our published figures and compare their trend line vs published data. Both are works that use our dataset, whether they went through terabytes of data, or used the trend line in the published paper.

When the papers use the dataset, the turbulence community typically use a citation to one of their own papers. But not all citations mean the papers are using the data. Sometimes they are using the data and results. On the other end of the spectrum it is just a citation and they have not even read the paper.

Ramanan feels that the impact of the work is not just the turbulence publications, But others that use the data in different forms.

We too have to write a "data management plan" for the INCITE proposal due on June 28th. We have not started writing that plan yet.

3.7 FUSION PARTICLE-IN-CELL – DAVID GREEN

3.7.1 Particle-In-Cell Workflow Overview

The PIC algorithm comes in both electrostatic and electromagnetic versions with most (if not all at large-scale) being an explicit time advance. There is work going on, in particular by L. Chacon et al. (http://w3fusion.ph.utexas.edu/~jift2012/uploads/11Chacon.pdf) on implicit time advance PIC schemes, but that is far from at-scale implementation as far as I know. Domain decomposition is achieved in one of two ways, either by particle set, or by spatial domain. In a particle set each processor tracks a fixed set of particles, and in the spatial DD particles are tracked only within a processors domain and handed off as they exit that domain. Most codes, as I understand it, decompose by spatial domain, but have to re-grid due to load balancing issues that can occur as the simulation progresses.

The general PIC algorithm goes like this ...

0. Load initial particle distribution
1. Scatter - Using the distribution of particles calculate the electric potential (or currents for electromagnetic) by depositing the charges on the grid.
2. Solve - for the electric field (and magnetic field for electromagnetic) from these above potentials via finite-difference or equivalent method.
3. Gather - forces on particles resulting from the above potentials.
4. Push - move particles for some delta t under the infulence of the above fields
5. Back to step 1.

As for the data usage / management requirements, a PIC simulation is characterized by the number of particles and the number of mesh points. Restart for a PIC code requires storage of all particle data, i.e., at least 6 variables per particle (3 space and 3 velocity coordinates, but usually it is a lot more, with something like 20 variables per particle), and all fields at each mesh point (3 electric and 3 magnetic field components plus potentials, etc so maybe another 20 variables per mesh point). For the larger simulations there may be of the order 1x10^9 particles and 1x10^8 mesh points. So you might expect

1x10^9 x 20 x 8 / (1024^3) = 150 GB for the particle data
1x10^8 x 20 x 8 / (1024^3) = 15 GB for the mesh data

assuming single (8 bits) precision.

Now that would be for restart only. You would often want to look at the time history of the fields, particle trajectories, etc. So there is typically some dump frequency at which a complete restart data set is dumped within a given simulation. For say 1000 data dumps you would end up with

1000 x (150 + 15) GB / (1024) = 160 TB per simulation.

Now I'm just estimating what might be a reasonable number of data dumps per simulation, but you can see how this gets very big, very fast, and that analyzing such large datasets requires parallel visualization capabilities like VisIt or the like.

Here is a good resource from NERSC

http://www.nersc.gov/research-and-development/benchmarking-and-workload-characterization/nersc-6-benchmarks/gtc/

3.7.2 Should data be preserved?

Institutions like Princeton have their own ; certainly some, the choice of what will be left to the PI and will need flexible guidelines of what to preserve (e.g., enough data so that someone could reproduce the pictures or the analysis in the paper) Don't want to just store.

3.7.3 Is Data Applicability Wide Or Narrow?

Benchmarking the codes; hard to do because of data; everything is subdivided and everyone is only interested in their and expert in.

3.7.4 Projected Future Data Requirements

If you need the time history, you need to see how things progress over time. Only the end result; the end may be a final 3d grid (several gb). you would have a set of fields for different runs. 100 different fields over several years.

4. CONCLUSION & JUDGMENTS OF NEED

4.1 FROM THE OLCF USER COMMUNITY PERSPECTIVE

DOIs provide a handy data science tool that users can use to

- Identify key data products of interest and value, and annotate them.
- Safely share data with their collaborators even before publishing the result in a scientific communication.
- Future data analyses can easily feed off of the data products, fostering a highly dynamic, and collaborative environment.
- Cite data products. The citations can even be tracked and could eventually contribute towards the user's h-index.
- List major data products as part of their two-page vita that is submitted along with their proposals. Funding agencies may begin to support this trend.
- Preserve data products for a longer-term, much beyond the expiration of their projects at the centers.
- Satisfy requirements from funding agencies on data management plans in terms of long-term preservation, sharing and dissemination of research results.

4.2 FROM THE OLCF PERSPECTIVE

From a center standpoint, DOIs offer the following benefits:

- It helps with research tracking and identifying the major results coming out of a project allocation on the center's resources.
- This can be extremely beneficial when it comes to reporting to sponsors.
- Since the DOIs also capture some basic metadata along with the index, it can help the center to answer questions on the disposition of the data, search and discover them.
- Finally, it also helps the center to cull the data holdings. Previously, users did not have a tool to identify what is important to them, which resulted in indiscriminately storing all intermediate snapshot data from scratch storage into archival storage. However, with DOIs, there is now a means to identify datasets of value, which may change this user behavior, resulting in manageable data sizes. This has ramifications to the provisioning of center storage resources.
- Provide tangible policies to users for long-term data preservation.
- Evolve to support "data-only" users through data science tools such as DOIs.
- DOI tools can be part of the data science repertoire that the center offers to users towards helping with their data management plans.

4.3 FROM A FUNDING SPONSOR PERSPECTIVE

From a sponsor standpoint:

- DOIs allow them to extract more value for the dollar spent. In addition to software tools, research artifacts, and papers, there is now a new entity, the citable data product.
- Added benefit of seeing data sharing flourish within the community, and more data analyses spawned from the data products.
- Better utilization of HPC center resources.
- Both users and centers that the sponsor funds now have rich tools for data management.

4.4 JUDGMENTS OF NEED

By offering to support a DOI infrastructure, a center such as OLCF is signing up for the following:

- Storage resources for the long-term preservation of data, typically on archival and staging storage.
- Network bandwidth during the retrieval of large data.
- Need for a data portal infrastructure for users to search and retrieve datasets based on DOIs. DOIs typically resolve to a landing page, which then provides means for data retrieval. The facility needs to maintain this portal and the associated infrastructure.
- Provide a quality assurance scheme to ensure the health of the datasets referred to by the DOIs. This is no different from providing the necessary quality assurance of the data products held in an HPSS archive.
- If a paid DOI registration authority is chosen, then the facility needs to maintain yearly subscriptions

.

5. REFERENCES

Anantharaj2012 Valentine Anantharaj, Ben Mayer, Feiyi Wang, Jim Hack, Daniel McKenna, and Rebecca Hartman-Baker Scientific Workflow and Support for High Resolution Global Climate Modeling at the Oak Ridge Leadership Computing Facility. Poster presented at European Geosciences Union General Assembly 2012. Vienna, Austria. April, 2012.

APPENDIX A. DATACITE METADATA

A.1. Proposed Fields

Table A1.1 provides a detailed description of the mandatory properties, together with their sub-properties, which *must* be supplied with any initial metadata submission to the managing agent for DataCite. In Table A1.2, the Recommended and Optional properties are described in detail. For an example of how to make a submission in XML format, please see the XML Examples provided on the DataCite Metadata Schema Repository1 website.

Throughout this document, a naming convention has been used for all properties and sub-properties as follows: properties begin with a capital letter, whereas sub-properties begin with a lower case letter. If the name is a compound of more than one word, subsequent words begin with capital letters.

The table uses shading to identify the combined Mandatory and Recommended "super set" of properties and sub-properties that enhance the prospect that the resource's metadata will be found, cited and linked.

The third column, Occurrence (Occ), indicates cardinality/quantity constraints for the properties as follows:

0-n = optional and repeatable
0-1 = optional, but not repeatable
1-n = required and repeatable
1 = required, but not repeatable

NOTE:
XML provides an xml:lang attribute2 that can be used on the properties `Title`, `Subject` and `Description`. This provides a way to describe the language used for the content of the specified properties. The schema provides a `Language` property to be used to describe the language of the resource.

1 http://schema.datacite.org/
2 Allowed values IETF BCP 47, ISO 639-1 language codes, e.g. en, de, fr

Table A1: Expanded DataCite Mandatory Properties

ID	***DataCite-Property***	***Occ***	***Definition***	***Allowed values, examples, other constraints***
1	Identifier	1	The Identifier is a unique string that identifies a resource.	DOI (Digital Object Identifier) registered by a DataCite member. Format should be "10.1234/foo"
1.1	identifierType	1	The type of the Identifier.	*Controlled List Value:* DOI
2	Creator	1-n	The main researchers involved in producing the data, or the authors of the publication, in priority order.	May be a corporate/institutional or personal name.
2.1	creatorName	1	The name of the creator.	Examples: Toru, Nozawa; Miller, Elizabeth The personal name format should be: family, given. Non-roman names may be transliterated according to the ALA-LC schemes3.
2.2	nameIdentifier	0-1	Uniquely identifies an individual or legal entity, according to various schemes.	The format is dependent upon scheme.
2.2.1	nameIdentifierScheme	1	The name of the name identifier scheme.	If nameIdentifier is used, nameIdentifierScheme is mandatory. Examples: ORCID4, ISNI5,
2.2.2	schemeURI	0-1	The URI of the name identifier scheme.	Examples : http://www.isni.org or http://www.orcid.org
3	Title	1-n	A name or title by which a resource is known.	Free text.
3.1	titleType	0-1	The type of Title.	*Controlled List Values:* AlternativeTitle Subtitle TranslatedTitle
4	Publisher	1	The name of the entity that holds, archives, publishes prints, distributes, releases, issues, or produces the resource. This property will be used to formulate the citation, so consider the prominence of the role.	Examples: World Data Center for Climate (WDCC); GeoForschungsZentrum Potsdam (GFZ); Geological Institute, University of Tokyo *** In the case of datasets, "publish" is understood to mean making the data available to the community of researchers.

3 http://www.loc.gov/catdir/cpso/roman.html
4 http://www.orcid.org/
5 http://www.isni.org/

ID	*DataCite-Property*	*Occ*	*Definition*	*Allowed values, examples, other constraints*
5	PublicationYear	1	The year when the data was or will be made publicly available.	YYYY *** If an embargo period has been in effect, use the date when the embargo period ends. If there is no standard publication year value, use the date that would be preferred from a citation perspective.

A.2. Expanded Properties

Table A2: Expanded DataCite Recommended and Optional Properties

ID	***DataCite-Property***	***Occ***	***Definition***	***Allowed values, examples, other constraints***
6	Subject	0-n	Subject, keyword, classification code, or key phrase describing the resource.	Free text.
6.1	subjectScheme	0-1	The name of the Subject scheme or classification code or authority if one is used.	Free text.
6.2	schemeURI	0-1	The URI of the subject identifier scheme.	Examples: http://id.loc.gov/authorities/subjects; http://dewey.info/
7	Contributor	0-n	The institution or person responsible for collecting, managing, distributing, or otherwise contributing to the development of the dataset.	
7.1	contributorType	1	The type of contributor of the resource.	If Contributor is used, then contributorType is mandatory. *Controlled List Values:* ContactPerson DataCollector DataManager Distributor Editor Funder HostingInstitution Producer ProjectLeader ProjectManager ProjectMember RegistrationAgency RegistrationAuthority RelatedPerson ResearchGroup RightsHolder Sponsor Supervisor WorkPackageLeader Other See appendix for definitions.

ID	*DataCite-Property*	*Occ*	*Definition*	*Allowed values, examples, other constraints*
7.2	contributorName	1	The name of the contributor.	If Contributor is used, then contributorName is mandatory. Examples: Patel, Emily; Doe, John The personal name format may be: family, given. Non-roman names should be transliterated according to the ALA-LC schemes6.
7.3	nameIdentifier	0-1	Uniquely identifies an individual or legal entity, according to various schemes.	The format is dependent upon scheme.
7.3.1	nameIdentifierScheme	1	The name of the name identifier scheme.	If nameIdentifier is used, nameIdentifierScheme is mandatory. Examples:ORCID7, ISNI8
7.3.2	schemeURI	0-1	The URI of the name identifier scheme.	Examples: http://www.isni.org http://www.orcid.org
8	Date	0-n	Different dates relevant to the work.	YYYY or YYYY-MM-DD or any other format described in W3CDTF.9 Use RKMS-ISO860110 standard for depicting date ranges Example: 2004-03-02/2005-06-02
8.1	dateType	1	The type of date.	If Date is used, dateType is mandatory. *Controlled List Values:* Accepted Available Copyrighted Collected Created Issued Submitted Updated Valid See appendix for definitions.
9	Language	0-1	The primary language of the resource.	Allowed values are taken from IETF BCP 47, ISO 639-1 language codes Examples: en, de, fr

6 http://www.loc.gov/catdir/cpso/roman.html
7 http://www.orcid.org/
8 http://www.isni.org/
9 http://www.w3.org/TR/NOTE-datetime
10 The standard is documented here: http://www.ukoln.ac.uk/metadata/dcmi/collection-RKMS-ISO8601/

ID	*DataCite-Property*	*Occ*	*Definition*	*Allowed values, examples, other constraints*
10	ResourceType	0-1	A description of the resource.	The format is open, but the preferred format is a single term of some detail so that a pair can be formed with the sub-property. *** Example: Image/Animation, where 'Image' is resourceTypeGeneral value and 'Animation' is ResourceType value.
10.1	resourceTypeGeneral	1	The general type of a resource.	If ResourceType is used, resourceTypeGeneral is mandatory. *Controlled List Values:* Audiovisual Collection Dataset Event Image InteractiveResource Model PhysicalObject Service Software Sound Text Workflow Other See appendix for definitions and examples.
11	AlternateIdentifier	0-n	An identifier or identifiers other than the primary Identifier applied to the resource being registered. This may be any alphanumeric string which is unique within its domain of issue. May be used for local identifiers.	Free text. Example: A local accession number: E-GEOD-34814
11.1	alternateIdentifierType	1	The type of the AlternateIdentifier.	If AlternateIdentifier is used, alternateIdentifierType is mandatory. Free text.
12	RelatedIdentifier	0-n	Identifiers of related resources. These must be globally unique identifiers.	Free text. *** Use this property to indicate subsets of properties, as appropriate.

ID	*DataCite-Property*	*Occ*	*Definition*	*Allowed values, examples, other constraints*
12.1	relatedIdentifierType	1	The type of the RelatedIdentifier	If RelatedIdentifier is used, relatedIdentifierType is mandatory. *Controlled List Values:* ARK DOI EAN13 EISSN Handle ISBN ISSN ISTC LISSN LSID PMID PURL UPC URL URN See appendix for full names and examples.
12.2	relationType	1	Description of the relationship of the resource being registered (A) and the related resource (B).	If RelatedIdentifier is used, relationType is mandatory. *Controlled List Values:* IsCitedBy Cites IsSupplementTo IsSupplementedBy IsContinuedBy Continues HasMetadata IsMetadataFor IsNewVersionOf IsPreviousVersionOf IsPartOf HasPart IsReferencedBy References IsDocumentedBy Documents isCompiledBy Compiles IsVariantFormOf IsOriginalFormOf IsIdenticalTo See appendix for definitions and examples.
12.3	relatedMetadataScheme	0-1	The name of the scheme.	Use only with this relation pair: (Has Metadata/ IsMetadataFor)

ID	*DataCite-Property*	*Occ*	*Definition*	*Allowed values, examples, other constraints*
12.4	schemeURI	0-1	The URI of the Related MetadataScheme.	Use only with this relation pair : (Has Metadata/ IsMetadataFor)
13	Size	0-n	Unstructured size information about the resource.	Free text. *** Examples: "15 pages", "6 MB"
14	Format	0-n	Technical format of the resource.	Free text. *** Use file extension or MIME type where possible, e.g., PDF, XML, MPG or application/pdf, text/xml, video/mpeg.
15	Version	0-1	The version number of the resource.	Suggested practice: track major_version.minor_version. Register a new identifier for a major version change. Individual stewards need to determine which are major vs. minor versions . May be used in conjunction with properties 11 and 12 (AlternateIdentifier and RelatedIdentifier) to indicate various information updates. May be used in conjunction with property 17 (Description) to indicate the nature and file/record range of version.
16	Rights	0-1	Any rights information for this resource.	Free text. *** Provide a rights management statement for the resource or reference a service providing such information. Include embargo information if applicable.
17	Description	0-n	All additional information that does not fit in any of the other categories. May be used for technical information.	The format is open *** It is a best practice to supply a description.

ID	*DataCite-Property*	*Occ*	*Definition*	*Allowed values, examples, other constraints*
17.1	descriptionType	1	The type of the Description.	If Description is used, descriptionType is mandatory. *Controlled List Values:* Abstract Methods SeriesInformation TableOfContents Other See appendix for definitions.
18	GeoLocation	0-n	Spatial region or named place where the data was gathered or about which the data is focused.	Repeat this property to indicate several different locations.
18.1	geoLocationPoint	0-1	A point location in space	A point contains a single latitude-longitude pair, separated by whitespace Detailed usage instructions[18] Example: <geoLocationPoint>31.233 -67.302 </geoLocationPoint>
18.2	geoLocationBox	0-1	The spatial limits of a place.	A box contains two white space separated latitude-longitude pairs, with each pair separated by whitespace. The first pair is the lower corner, the second is the upper corner. Detailed usage instructions11 Example: <geoLocationBox>42.893-71.032 41.090-68.211</geoLocationBox>
18.3	geoLocationPlace	0-1	Description of a geographic location	Free text. Use to describe a geographic location.

11 Use WGS 84 (World Geodetic System) coordinates. Use only decimal numbers for coordinates. Longitudes are -180 to 180 (0 is Greenwich, negative numbers are west, positive numbers are east), Latitudes are -90 to 90 (0 is the equator; negative numbers are south, positive numbers north.

A.3. ContributorType Values

Table A3: Description of contributorType

Option	***Description***	***Usage Notes***
ContactPerson	Person with knowledge of how to access, troubleshoot, or otherwise field issues related to the resource	May also be "Point of Contact" in organization that controls access to the resource, if that organization is different from Publisher, Distributor, Data Manager
DataCollector	Person/institution responsible for finding, gathering/collecting data under the guidelines of the author(s) or Principal Investigator (PI)	May also use when crediting survey conductors, interviewers, event or condition observers, person responsible for monitoring key instrument data.
DataManager	Person (or organization with a staff of data managers, such as a data center) responsible for maintaining the finished resource.	The work done by this person or organization ensures that the resource is periodically "refreshed" in terms of software/hardware support, is kept available or is protected from unauthorized access, is stored in accordance with industry standards, and is handled in accordance with the records management requirements applicable to it.
Distributor	Institution tasked with responsibility to generate/disseminate copies of the resource in either electronic or print form.	Works stored in more than one archive/repository may credit each as a distributor.
Editor	A person who oversees the details related to the publication format of the resource.	
Funder	Institution that provided financial support for the development of the resource.	Recommended for discovery. Includes organizations that provide funding via regular budget allocations, through grants or awards
HostingInstitution	Typically, the organization allowing the resource to be available on the Internet through the provision of its hardware/software/operating support.	May also be used for an organization that stores the data offline. Often a data center (if that data center is not the "publisher" of the resource.
Producer	Typically a person or organization responsible for the artistry and form of a media product.	In the data industry, this may be a company "producing" DVDs that package data for future dissemination by a distributor.
ProjectLeader	Person officially designated as head of project team or sub-	The Project Leader is not "removed' from the work that

Option	Description	Usage Notes
	project team instrumental in the work necessary to development of the resource.	resulted in the resource; he or she remains intimately involved throughout the life of the particular project team.
ProjectManager	Person officially designated as manager of a project. Project may consist of one or many project teams and sub-teams.	The manager of a Project normally has more administrative responsibility than actual work involvement.
ProjectMember	Person on the membership list of a designated project/project team	This vocabulary may or may not indicate the quality, quantity, or substance of the person's involvement.
RegistrationAgency	Institution/organization officially appointed by a Registration Authority to handle specific tasks within a defined area of responsibility	DataCite is a Registration Agency for the International DOI Foundation (IDF). One of DataCite's tasks is to assign DOI prefixes to the allocating agents who then assign the full, specific character string to data clients, provide metadata back to the DataCite registry, etc.
RegistrationAuthority	A standards-setting body from which Registration Agencies obtain official recognition and guidance.	The IDF serves as the Registration Authority for the International Standards Organization (ISO) in the area/domain of Digital Object Identifiers.
RelatedPerson	A person without a specifically defined role in the development of the resource, but who is someone the author wishes to recognize.	This person could be an author's intellectual mentor, a person providing intellectual leadership in the discipline or subject domain, etc.
ResearchGroup	Typically refers to a group of individuals with a lab, department, or division; the group has a particular, defined focus of activity.	May operate at a narrower level of scope; may or may not hold less administrative responsibility than a project team.
RightsHolder	Person or institution owning or managing property rights, including intellectual property rights over the resource.	
Sponsor	Person or organization that issued a contract or under the auspices of which a work has been written, printed, published, developed, etc.	Includes organizations that provide in-kind support, through donation, provision of people or a facility or instrumentation necessary for the development of the resource, etc.
Supervisor	Designated administrator over one or more groups/teams working to produce a resource or over one or more steps of a development process.	
WorkPackageLeader	A Work Package is a recognized	

Option	*Description*	*Usage Notes*
	data product, not all of which is included in publication. The package, instead, may include notes, discarded documents, etc. The Work Package Leader is responsible for ensuring the comprehensive contents, versioning, and availability of the Work Package during the development of the resource.	
Other	Any person or institution making a significant contribution to the development and/or maintenance of the resource, but whose contribution does not "fit" other controlled vocabulary for ContributorType.	Could be a photographer, artist, or writer whose contribution helped to publicize the resource (as opposed to creating it), a reviewer of the resource, someone providing administrative services to the author (such as depositing updates into an online repository, analysing usage, etc.), or one of many other roles.

A.4. DataType Values

Table A.4: Description of dateType

Option	***Description***	***Usage Notes***
Accepted	The date that the publisher accepted the resource into their system.	To indicate the start of an embargo period, use Submitted or Accepted, as appropriate.
Available	The date the resource is made publicly available. May be a range.	To indicate the end of an embargo period, use Available.
Copyrighted	The specific, documented date at which the resource receives a copyrighted status, if applicable.	
Collected	The date or date range in which the resource content was collected.	To indicate precise or particular timeframes in which research was conducted.
Created	The date the resource itself was put together; this could be a date range or a single date for a final component, e.g., the finalised file with all of the data.	Recommended for discovery.
Issued	The date that the resource is published or distributed e.g. to a data center	
Submitted	The date the creator submits the resource to the publisher. This could be different from Accepted if the publisher then applies a selection process.	Recommended for discovery. To indicate the start of an embargo period, use Submitted or Accepted, as appropriate.
Updated	The date of the last update to the resource, when the resource is being added to. May be a range.	
Valid	The date or date range during which the dataset or resources are accurate. May be a range.	

A.5. ResourceType Values

Table A.5: Description of resourceTypeGeneral

Option	***Description12***	***Examples and Usage Notes***	***Suggested Dublin Core Mapping***
Audiovisual	A series of visual representations imparting an impression of motion when shown in succession. May or may not include sound.	May be used for films, video, etc, Ex: http://data.datacite.org/10.7794/USR3225/ARC.F/0000001	MovingImage
Collection	An aggregation of resources of various types. If a collection exists of a single type, use the single type to describe it.	A collection of various files making up a report. Ex: http://data.datacite.org/10.5284/1001038	Collection
Dataset	Data encoded in a defined structure.	Data file or files, Ex: http://data.datacite.org/10.4231/D39Z90B9T	Dataset
Event	A non-persistent, time-based occurrence	Materials associated with an event in time. Ex: http://data.datacite.org/10.7269/P3RN35SZ	Event
~~Film~~	**DEROGATED** A series of visual representations imparting an impression of motion when shown in succession. May or may not include sound.	**DEROGATED; use audiovisual instead.**	**DEROGATED** MovingImage
Image	A visual representation other than text.	Digitized or born digital photographs. Ex: http://data.datacite.org/10.5060/D4RN35SD	Image, StillImage
InteractiveResource	A resource requiring interaction from the user to be understood, executed, or experienced	Training modules, files that require use of a viewer, or query/response portals. Ex: http://data.datacite.org/10.7269/P3TB14TR	InteractiveResource

12Where there is direct correspondence with the Dublin Core Metadata, DataCite definitions have borrowed liberally from the DCMI definitions. See: http://dublincore.org/documents/dcmi-terms/#terms-DCMIType

Option	***Description12***	***Examples and Usage Notes***	***Suggested Dublin Core Mapping***
Model	An abstraction of the real thing, i.e. some generalisation and interpretation. A symbolic representation.	Modelled descriptions of, for example, different aspects of languages or a molecular biology reaction chain.	N/A
PhysicalObject	An inanimate, three-dimensional object or substance.	Artifacts, specimens Ex: http://data.datacite.org/10.7299/X78052RB	PhysicalObject
Service	A system that provides one or more functions of value to the end-user.	Data management service, authentication service,	Service
Software	A computer program in source or compiled form.	Software supporting research. Ex: http://data.datacite.org/10.5524/100046	Software
Sound	A resource primarily intended to be heard.	Audio recording. Ex: http://data.datacite.org/10.4231/3DZVR.1288284313124	Sound
Text	A resource consisting primarily of words for reading.	Grey literature, lab notes, accompanying materials. Ex: http://data.datacite.org/10.5682/9786065914018	Text
Workflow	A structured series of steps which can be executed to produce a final outcome, allowing users a means to specify and enact their work in a more reproducible manner.	Computational workflows involving sequential operations made on data by wrapped software and may be specified in a format belonging to a workflow management system, such as Taverna (http://www.taverna.org.uk/).	N/A
Other	If selected, supply a value for ResourceType.		

A.6. RelationType Values

Table A.6: Description of relationType

Option	***Definition***	***Example and Usage Notes***
IsCitedBy	indicates that B includes A in a citation	Recommended for discovery. <relatedIdentifier relatedIdentifierType="DOI"relationType="IsCitedBy">10.4232/10.ASEAS-5.2-1 </relatedIdentifier>
Cites	indicates that A includes B in a citation	Recommended for discovery. <relatedIdentifier relatedIdentifierType="ISBN" relationType="Cites“>0761964312 </relatedIdentifier>
IsSupplementTo	indicates that A is a supplement to B	Recommended for discovery. <relatedIdentifier relatedIdentifierType="URN" relationType="IsSupplementTo">http://nbn-resolving.de/urn:nbn:de:0168-ssoar-13172 </relatedIdentifier>
IsSupplementedBy	indicates that B is a supplement to A	Recommended for discovery. <relatedIdentifier relatedIdentifierType="PMID" relationType="IsSupplementedBy">16911322/ </relatedIdentifier>
IsContinuedBy	indicates A is continued by the work B	<relatedIdentifier relatedIdentifierType="URN" relationType="IsContinuedBy">http://nbn-resolving.de/urn:nbn:de:bsz:21-opus-4967 </relatedIdentifier>
Continues	indicates A is a continuation of the work B	<relatedIdentifier relatedIdentifierType="URN" relationType="Continues">http://nbn-resolving.de/urn:nbn:de:bsz:21-opus-4966 </relatedIdentifier>
HasMetadata	indicates A is relates to an external file of additional metadata B	<relatedIdentifier relatedIdentifierType="DOI" relationType="HasMetadata">10.1234/567890 relatedMetadataScheme="DDI-L" schemeURI="http://info1.gesis.org/DDI/3_1/instance.xsd"</relatedIdentifier>
IsMetadataFor	indicates A is additional metadata for a work or resource B	<relatedIdentifier relatedIdentifierType="DOI" relationType="IsMetadataFor">10.1234/567891 relatedMetadataScheme="DDI-L" schemeURI="http://info1.gesis.org/DDI/3_1/instance.xsd"</relatedIdentifier>

Option	*Definition*	*Example and Usage Notes*
IsNewVersionOf	indicates A is a new edition of B, where the new edition has been modified or updated	<relatedIdentifier relatedIdentifierType="DOI" relationType="IsNewVersionOf">10.5438/0005 </relatedIdentifier>
IsPreviousVersion Of	indicates A is a previous edition of B	<relatedIdentifier relatedIdentifierType="DOI" relationType="IsPreviousVersionOf">10.5438/0007 </relatedIdentifier>
IsPartOf	indicates A is a portion of B; may be used for elements of a series	Recommended for discovery. <relatedIdentifier relatedIdentifierType="ISBN" relationType="IsPartOf">0-486-27557-4 </relatedIdentifier>
HasPart	indicates A includes the part B	Recommended for discovery. <relatedIdentifier relatedIdentifierType="DOI" relationType="HasPart">10.1234/7894 </relatedIdentifier>
IsReferencedBy	indicates A is used as a source of information by B	<relatedIdentifier relatedIdentifierType="URL" relationType="IsReferencedBy">http://www.testpubl.de </relatedIdentifier>
References	indicates B is used as a source of information for A	<relatedIdentifier relatedIdentifierType="URN" relationType="References">http://nbn-resolving.de/urn:nbn:de:bsz:21-opus-963</relatedIdentifier>
IsDocumentedBy	indicates B is documentation about/explaining A)	<relatedIdentifier relatedIdentifierType="URL" relationType="IsDocumentedBy">http://tobias-lib.uni-tuebingen.de/volltexte/2000/96/ </relatedIdentifier>
Documents	indicates A is documentation about/explaining B	<relatedIdentifier relatedIdentifierType="DOI" relationType="Documents">10.1234/7836 </relatedIdentifier>
isCompiledBy	indicates B is used to compile or create A	<relatedIdentifier relatedIdentifierType="URL" relationType="isCompiledBy">http://d-nb.info/gnd/4513749-3 </relatedIdentifier>
Compiles	indicates B is the result of a compile	<relatedIdentifier relatedIdentifierType="URN" relationType="Compiles">http://nbn-resolving.de/urn:nbn:de:bsz:21-opus-963

Option	*Definition*	*Example and Usage Notes*
	or creation event using A	</relatedIdentifier>
IsVariantFormOf	indicates A is a variant or different form of B, e.g. calculated or calibrated form or different packaging	<relatedIdentifier relatedIdentifierType="DOI" relationType="IsVariantFormOf">10.1234/8675 </relatedIdentifier>
IsOriginalFormOf	indicates A is the original form of B	<relatedIdentifier relatedIdentifierType="DOI" relationType="IsOriginalFormOf">10.1234/9035 </relatedIdentifier>
IsIdenticalTo	indicates that A is identical to B, for use when there is a need to register two separate instances of the same resource	<relatedIdentifier relatedIdentifierType="URL" relationType="IsIdenticalTo">http://oac.cdlib.org/findaid/ark:/13030/c8r78fzq </relatedIdentifier>

A.7. DescriptionType Values

Table A.7: Description of contributorType

Option	***Definition***	***Usage Notes***
Abstract	A brief description of the resource and the context in which the resource was created.	Recommended for discovery. Use " " to indicate a line break for improved rendering of multiple paragraphs, but otherwise no html markup. Example: http://data.datacite.org/10.1594/PANGAEA.771774
Methods	The technology methodology employed for the study or research.	Recommended for discovery. For example, see section "Sampling, Processing and Quality Control Methods" in the following dataset record: http://knb.ecoinformatics.org/knb/metacat?action=read&qformat=knb&sessionid=0&docid=knb-lter-gce.275.16.
SeriesInformation	Information about a repeating series, such as volume, issue, number.	For use with grey literature. If providing an ISSN, use property 12 (RelatedIdentifier), relatedIdentifierType=ISSN. For dataset series, use property 12(RelatedIdentifier) and describe the relationships with isPartOf or HasPart. Example: http://data.datacite.org/10.4229/23RDEUPVSEC2008-5CO.8.3
TableOfContents	A listing of the Table of Contents.	Use " " to indicate a line break for improved rendering of multiple paragraphs, but otherwise no html markup. Example: http://data.datacite.org/10.5678/LCRS/FOR816.CIT.1031
Other	Other description information that does not fit into an existing category.	Use for any other description type.

APPENDIX B. DOE ANNOUNCEMENT NOTICE 241.6

B.1 Background

B.1.1 Announcement Notice 241.6

Announcement Notice (AN) 241.6 provides to the U.S. Department of Energy (DOE) Office of Scientific and Technical Information (OSTI) metadata needed to identify/announce publicly available datasets resulting from work funded by the DOE or performed in DOE facilities. The submitted information also allows OSTI to assign Digital Object Identifiers (DOIs) to datasets and register them with DataCite as a service to researchers. This value-added step facilitates visibility, helps ensure long-term preservation, and supports better linkage between DOE's published research results and the underlying data. The DOI assigned to each announced dataset is included in the XML response to each successful submittal. The primary contact identified in the metadata will also be notified by an automated email of successfully submitted records and the DOIs assigned.

The datasets themselves must be publicly available and maintained at a DOE site, DOE Data Center, or other publicly accessible location, such as an open repository. Submitting sites agree to ensure data persistence, which includes storing and managing data such that access and usability are provided indefinitely. The URL associated with each DOI should point to an HTML "landing page" that provides context for the dataset or to a notice page ("tombstone page") for data that has had to be retired. Datasets are not uploaded or stored at OSTI.

.The AN 241.6 web service is the easiest way to submit multiple records on a regular basis. However, an interface for manual entry of one AN 241.6 record at a time is available on E-Link, if desired.

B.1.2 OSTI's Web Service for AN 241.6

OSTI's Web Service for AN 241.6 provides an easy-to-use mechanism to submit new Announcement Notices, edit existing Announcement Notices, and retrieve metadata for Scientific and Technical Information (STI) Announcement Notices. A user manipulates a record by performing HTTP operations on the web service URL and providing XML metadata. The POST command/request allows a submitter to create or modify AN 241.6 records. The GET command/request allows a submitter to view records.

Submitters must coordinate with OSTI and submit records to a test environment before POSTing their first file. The test URL is https://www.osti.gov/elinktest/2416api. OSTI's production web service for announcing datasets is available at https://www.osti.gov/elink/2416api.

All transactions with the web service require authentication through an active E-Link user account. For details or to request access to E-Link and obtain an active account, please visit https://www.osti.gov/elink/register.jsp.

B.2.0 Using the STI Announcement Web Service

B.2.1 Authentication

Each request requires authentication through an active E-Link user account. OSTI's STI Announcement Web Services supports HTTP Basic authentication over SSL. With this method, the client connects an

HTTPS URL (e.g., https://www.osti.gov/elink/2416api). The POST and GET verb commands will pass along the standard Authentication HTTP header (base64 encoding).

If authentication is successful, the input body of the request is read as an XML document, parsed, and submitted appropriately. HTML status codes (200=OK, 401=Unauthorized, 500=System error, etc.) and an XML response, which includes certain metadata fields, are returned at the end of processing each request. Additional authentication steps may be taken prior to editing a metadata record.

B.2.2 POST Command: Creating and Modifying Records

When submitting data via the 2416 web service, a secure connection is created, and the login/authentication information and the metadata records are streamed through the connection. The OSTI web service will generate a response indicating success or failure for each records. An error message will be included for each failed record, as well. Multiple records can be submitted in a single submission.

The following sample of Java code illustrates a submission with the POST command. The sample name "testPost" can, of course, be changed to any name desired.

```
protected static  boolean testPost( ) throws IOException {
      HttpURLConnection c = null;
      DataOutputStream out = null;
      InputStreamReader in = null;
      BufferedReader br = null;

      try {

            /** build the URL to connect to */
            StringBuilder url = new StringBuilder();
            // urlPart is defined elsewhere - it is the "https://www.osti....."
            url.append(urlPart);
            URL u = new URL(url.toString());
            c = (HttpURLConnection) u.openConnection();
            c.setRequestMethod("POST");
            c.setRequestProperty("Accept", "application/xml");
            c.setRequestProperty("Content-Type", "application/xml;charset=UTF-8");

            /*** add login authentication
                usr_passwd is a string defined elsewhere. It is USERNAME + ": " + PASSWORD for the
                elink user account
            ***/
            String auth = org.apache.commons.codec.binary.Base64.encodeBase64URLSafeString(
(usr_passwd).getBytes() );
            c.setRequestProperty("Authorization", "Basic " + auth);

          /** write out the metadata stream */
          c.setDoOutput(true);
          out = new DataOutputStream(c.getOutputStream());

          /** make a string that contains the record(s) in XML format
              YOUR SITE DATA GOES IN THIS STRING
          */
          String rec = [SITE DATA];

          /** add the rec string to the output stream **/
          byte[] buf = (null==rec) ? "".getBytes("UTF-8") : rec.getBytes("UTF-8");
          out.write(buf, 0, buf.length);
          out.close(); out = null;

          /** open the connection **/
          c.connect();
```

```
      /** get the response and appropriate response information */
      int responseCode = c.getResponseCode();
      if (responseCode>=400) in = new InputStreamReader(c.getErrorStream());
      else                   in = new InputStreamReader(c.getInputStream());

      br = new BufferedReader(in);
      StringWriter writer = new StringWriter();
      char[] buffer = new char[1024];
      int n=0;
      while ((n=br.read(buffer))!=-1) {
         writer.write(buffer, 0, n);
      }
      writer.close();

      // see what the test results are from the API servlet
      StringBuffer outBuf = writer.getBuffer();
      System.out.print("Response returned " + outBuf.toString());

      /** log any errors */
      if (responseCode!=200) {
         log.error("OSTI ID: " + m.getOstiId() + " failed to post new DOI, error code=" + responseCode);
         log.error("Message: " + writer.toString());
      }
      /** finished */
      return (200==responseCode);
   } finally {
      try {
         if (br!=null) br.close(); br = null;
         if (in!=null) in.close(); in = null;
         if (out!=null) out.close(); out = null;
      } catch ( Exception e ) {
         log.error("URL Close Error: " + e.getMessage());
      }
   }
}
```

B.2.3 GET Command: Requesting Records

Metadata can be retrieved for records previously submitted from your site by using a GET request and supplying the osti_id argument on the command line. Authentication is required and is handled in the same fashion as a Create/Modify command. Metadata is returned as XML.

The following sample of Java code illustrates a GET request. The sample name “testGet” can, of course, be changed to any name desired.

```
protected static boolean testGet( ) throws IOException {

   HttpURLConnection c = null;
   DataOutputStream out = null;
   InputStreamReader in = null;
   BufferedReader br = null;

   try {
      /** build the URL to connect to */
      StringBuilder url = new StringBuilder();
      // urlPart is defined elsewhere - it is the "https://www.osti...../2416API"
      url.append(urlPart);
      /** append the osti id parameter**/
      url.append("?osti_id=1001628");
      URL u = new URL(url.toString());
      c = (HttpURLConnection) u.openConnection();
      c.setRequestMethod("GET");
      c.setRequestProperty("Accept", "application/xml");
      c.setRequestProperty("Content-Type", "application/xml;charset=UTF-8");
```

```
            /** * add login authentication
                usr_passwd is a string defined elsewhere. It is  USERNAME + ": " + PASSWORD for the
                elink user account
            * * */
            String auth = org.apache.commons.codec.binary.Base64.encodeBase64URLSafeString(
(usr_passwd).getBytes() );
            c.setRequestProperty("Authorization", "Basic " + auth);
            c.connect();

            /** get the response and appropriate response information */
            int responseCode = c.getResponseCode();

            if (responseCode>=400) in = new InputStreamReader(c.getErrorStream());
            else                   in = new InputStreamReader(c.getInputStream());

            br = new BufferedReader(in);
            StringWriter writer = new StringWriter();
            char[] buffer = new char[1024];
            int n=0;
            while ((n=br.read(buffer))!=-1) {
               writer.write(buffer, 0, n);
            }
            writer.close();

            /** log any errors */
            if (responseCode!=200) {
               log.error("OSTI ID: 10011628; error code=" + responseCode);
               log.error("Message: " + writer.toString());
            }
           /** finished */


         } finally {
            try {
               if (br!=null) br.close(); br = null;
               if (in!=null) in.close(); in = null;
               if (out!=null) out.close(); out = null;
            } catch ( Exception e ) {
               log.error("URL Close Error: " + e.getMessage());
            }
         }
      }
```

B.3.0 AN 241.6 Web Service Metadata

B.3.1 AN 241.6 Required Fields

The following is a list of the required fields for the AN 241.6 STI Announcement Web Service. Required fields are designated by an asterisk (*). Records without required fields will fail to load into E-Link for processing and DOI registration. Only 15 metadata and/or administrative fields are required. The other fields available for your use are optional, though some, such as the Abstract/Description, are highly encouraged.

Note that the OSTI ID is a required field for all POST requests where the intent is to edit or update records. The GET request must also include the OSTI ID and will allow retrieval of a record previously submitted by your site.

REQUIRED

	Field Name	XML Tag Name	Additional Information
1	OSTI ID+	<osti_id>	Note that the OSTI ID is required in all requests intended to edit or update records.

			When POSTing new records to OSTI, no <osti_id> tag is needed in the XML. E-Link automatically assigns an OSTI ID to each record successfully submitted; you will receive it in the XML response returned to your site by the OSTI webservice.
2	Site Code*	N/A	Automatically determined by the authenticated E-Link User account
3	Dataset Type*	<dataset_type/>	Dataset Type refers to the main content of the dataset. Only one value is allowed. Use the two-letter code shown below:
		Code	**Definition**
		AS	Animations/Simulations
		GD	Genome Data - Information that is numeric or alpha-numeric in nature (such as gene sequences) or that is a specialized mix of text and non-text information conveying results of genetics/genome research
		IM	Interactive Data Map(s) – A non-static interface and the GIS data and/or shape files that generate it.
		ND	Data primarily expressed with numbers; other content is secondary and supporting.
		IP	Still Images or Photos - A collection of images or photographs produced by a scientific instrument or that convey scientific results of experiments. Scientific images that might constitute a data set could be images of cells or molecules that are typically taken with electron microscopes, 3-D structures of proteins or nanomaterials, images captured during an accelerator run, images from astronomy, etc.
4	Dataset Title*	<title/>	
5	Creator(s)/ Principal Investigator(s)*	<creators/>	Format is Last Name, First Name, MI Separate multiple authors with a semi-colon followed by a space.
6	Dataset Product Number(s)*	<product_nos/>	The most important identifying numbers given to the dataset by the host or originating organization. Separate multiple values with a semi-colon followed by a space. 'None" is an acceptable value when necessary.
7	DOE Contract Number(s)*	<contract_nos/>	Use the format of the contract "as is," but please leave off any preceding "DE". If multiple contract and/or grant numbers apply, separate with a semi-colon followed by a space.
8	Originating Research Organization*	<originating_research_org/>	Use the spelled-out text exactly as shown in the Originating Research Organization Authority at https://www.osti.gov/elink/authorities.jsp

			If work for this product was done at more than one research organization, multiple values may be listed; they should be separated by a semicolon and a space. The primary DOE organization should be listed first, followed by any others. If non-DOE orgs are included, input the spelled-out, full name of the organization.
9	Publication/Issue Date*	<publication_date/>	Use one of these three Publication Date formats: • mm/dd/yyyy • yyyy • yyyy Month
10	Language*	<language/>	Up to75 characters; use format from OSTI's Language Authority; e.g. English Authority values are available at https://www.osti.gov/elink/authorities.jsp
11	Country of Origin/ Publication*	<country/>	Use two character code from OSTI's Country Code Authority; e.g. US Authority values are available at https://www.osti.gov/elink/authorities.jsp
17	Sponsoring Organization(s)*	<sponsor_org/>	Use the spelled-out text as shown in the Sponsoring Organization Authority at https://www.osti.gov/elink/authorities.jsp If funding for this product was provided from more than one organization, multiple values may be listed; they should be separated by a semicolon and a space. The primary DOE sponsor should be listed first, followed by any others. If any of the others are not included in the Sponsor Organization Authority (non-DOE organizations, for example), please include the spelled-out, full name of the other sponsoring organization.
12	Site URL*	<site_url/>	OSTI can not accept, store, or post datasets. Datasets must be publicly available, and the submitted metadata must include a valid URL. The URL should link to an html "landing page" for the dataset.
13	Contact Name and Position*	<contact_name/>	Admin info only; it will not be displayed in public databases.
14	Contact Organization*	<contact_org/>	Admin info only; it will not be displayed in public databases.
15	Contact E-mail*	<contact_email/>	Admin info only; it will not be displayed in public databases.

B.3.2 AN 241.6 Optional Fields

The following is a list of optional fields for the AN 241.6 STI Announcement Web Service. Inclusion of some of these fields, such as the Abstract/Description, is highly encouraged, however.

OPTIONAL

1	Creator(s)/PI Email Address(es)	<creators_emails/>	Admin info only; it will not be displayed in public databases.
2	Related Resource	<related_resource/>	This is a place to provide the bibliographic info on the key paper(s) that the dataset supports.
3	Availability	<availability/>	Normally used to provide the name of an organization, a division within a lab, a specific employee's title, etc. to which a request for further information may be made.
4	Contributor Organizations	<contributor_ organizations/>	Provide the names of any organizations that have significantly contributed to the gathering, formatting, analysis, etc. of the dataset. These are organizations that would not otherwise be credited because they will not be listed in the Originating Research/Submitting Organization field, or in the Sponsoring Organization field. Separate multiple entries with a semicolon and a space.
5	Other Identifying Numbers(s)	<other_identifying_nos/>	Any other numbers that users might wish to retrieve on or need to recognize. If there are multiple values in this field, separate them with a semicolon followed by a space.
6	Subject Categories	<subject_categories_ code/>	Use the complete value (numerical code and spelled-out category title) as shown in the Subject Category Authority at https://www.osti.gov/elink/authorities.jsp. As many multiples as needed are allowed in this tag set; separate them with a semicolon and a space. List the primary subject category first.
7	Keywords	<keywords/>	
8	Description/ Abstract	<description/> 4000 character limit	Provide a clear, concise summary of the content of the dataset, as well as specialized parameters that describe the data. Specialized parameters may include a date range during which information was taken (such as May, 01 2002 - December 31, 2002), geographic information (such as a specific state, region, country, latitude and longitude, etc.), information such as well depth ranges, temperature ranges, etc.
9	DOI	<doi/>	Provide the DOI if one has been assigned prior to the dataset being announced to OSTI.
10	Dataset's File Extension	<file_extension/>	Some common file extensions are .txt, .csv, .ps, etc.
11	Software needed to utilize dataset	<software_needed/>	Specialized software tools are often developed to allow a user to manipulate data in various ways. If these tools are available for the user but do not have to be used with the data, they do not need to be listed. However, if there is a piece of software without which a user cannot open, see, or use the dataset, that software should be noted in this field
12	Dataset Size	<dataset_size/>	Optional. Indicate approximate size in number of files, in megabytes, or in other ways appropriate for the dataset's content.
13	Contact Phone	<contact_phone/>	Admin info only; it will not be displayed in public databases.

B.3.3 Examples of XML 241.6 Metadata Records

Here is an example of a 241.6 dataset record as it would come to OSTI's 241.6 web service. Immediately following it is what the xml response from web service back to the submitting service would be in case of a successful submission and what you would see if the submission failed. POST SUBMISSION SAMPLE:

<?xml version="1.0" encoding="UTF-8"?>
<records>
<record>
<osti_id></osti_id>
<dataset_type>ND</dataset_type>
<title>ARM Climate Modeling Best Estimate Lamont, OK (ARMBE-CLDRAD SGPC1)</title>
<creators>Renata McCoy; Shaocheng Xie;</creators>
<creators_emails></creators_emails>
<related_resource></related_resource>
<product_nos>none</product_nos>
<contract_nos>AC05-00OR22725</contract_nos>
<other_identifying_numbers>sgpC1amrbe-cldrd-v3</other_identifying_numbers>
<availability></availability>
<contributor_organizations>Pacific Northwest National Laboratory (PNNL); Brookhaven National Laboratory (BNL); Argonne National Laboratory (ANL); Oak Ridge National Laboratory (ORNL)</contributor_organizations>
<publication_date>05/14/2012</publication_date>
<language>English</language>
<country>US</country>
<sponsor_org>USDOE Office of Science (SC), Biological and Environmental Research (BER)</sponsor_org>
<subject_categories_code>54 Environmental Sciences</subject_categories_code>
<keywords>Cloud fraction profiles; Total, high, middle, and low clouds; Liquid water path and precipitable water vapor; Surface radiative fluxes; TOA radiative fluxes</keywords>
<description>The ARM CMBE-ATM [Xie, McCoy, Klein et al.] data file contains a best estimate of several selected atmospheric quantities from ACRF observations and NWP analysis data.</description>
<site_url>http://iop.archive.arm.gov/arm-iop/0showcase-data/cmbe/cmbe/sgpC1/cmbe-cldrad/</site_url>
<doi></doi>
<file_extension>cdf</file_extension>
<software_needed></software_needed>
<dataset_size>12544 KB</dataset_size>
<contact_name> ARM Archive User Services</contact_name>
<contact_org> ORNL</contact_org>
<contact_email> armarchive@ornl.gov</contact_email>
<contact_phone> 888-276-3282</contact_phone>
</record>
</records>

POST SUCCESSFUL – SAMPLE RETURN MESSAGE

<?xml version="1.0" encoding="UTF-8"?>
<records>
<record>
<osti_id>1035366</osti_id>
<product_nos>none</product_nos>
<title>ARM Climate Modeling Best Estimate Lamont, OK (ARMBE-CLDRAD SGPC1)</title>
<contract_nos>AC05-00OR22725</contract_nos>
<doi>http://dx.doi.org/10.5439/1035366</doi>
<status>SUCCESS</status>
<status_message></status_message>
</record>
</records>

POST FAILURE – SAMPLE RETURN MESSAGE

<?xml version="1.0" encoding="UTF-8"?>
<records>
<record>
<osti_id>0</osti_id>
<product_nos>none</product_nos>
<title>ARM Climate Modeling Best Estimate Lamont, OK (ARMBE-CLDRAD SGPC1)</title>
<contract_nos>AC05-00OR22725</contract_nos>
<doi></doi>

```
<status>FAILURE</status>
<status_message>Data too long, maximum number of characters for dataset type is 2</status_message>
</record>
</records>
```

The “FAILURE” tells the submitting organization that no record was loaded into E-Link in this instance; that’s why the OSTI ID number is 0. There’s no DOI number assigned to the dataset because DOI assignment happens during processing…which never took place. The status message identifies the error that made the POST submission fail. In this case, if the submitted record had had a three letter code in the dataset_type field instead of the correct 2 character ND, this would have been the error.

It is the submitting site’s responsibility to review the returned messages, correct any errors, and resubmit the failed records.

APPENDIX C. THE HOLDREN MEMO

Official Office of Science and Technology Policy Response to Require free access over the Internet to scientific journal articles arising from taxpayer-funded research.

Increasing Public Access to the Results of Scientific Research

By Dr. John Holdren

Thank you for your participation in the We the People platform. The Obama Administration agrees that citizens deserve easy access to the results of research their tax dollars have paid for. As you may know, the Office of Science and Technology Policy has been looking into this issue for some time and has reached out to the public on two occasions for input on the question of how best to achieve this goal of democratizing the results of federally-funded research. Your petition has been important to our discussions of this issue.

The logic behind enhanced public access is plain. We know that scientific research supported by the Federal Government spurs scientific breakthroughs and economic advances when research results are made available to innovators. Policies that mobilize these intellectual assets for re-use through broader access can accelerate scientific breakthroughs, increase innovation, and promote economic growth. That's why the Obama Administration is committed to ensuring that the results of federally-funded scientific research are made available to and useful for the public, industry, and the scientific community.

Moreover, this research was funded by taxpayer dollars. Americans should have easy access to the results of research they help support.

To that end, I have issued a memorandum today (.pdf) to Federal agencies that directs those with more than $100 million in research and development expenditures to develop plans to make the results of federally-funded research publically available free of charge within 12 months after original publication. As you pointed out, the public access policy adopted by the National Institutes of Health has been a great success. And while this new policy call does not insist that every agency copy the NIH approach exactly, it does ensure that similar policies will appear across government.

As I mentioned, these policies were developed carefully through extensive public consultation. We wanted to strike the balance between the extraordinary public benefit of increasing public access to the results of federally-funded scientific research and the need to ensure that the valuable contributions that the scientific publishing industry provides are not lost. This policy reflects that balance, and it also provides the flexibility to make changes in the future based on experience and evidence. For example, agencies have been asked to use a 12-month embargo period as a guide for developing their policies, but also to provide a mechanism for stakeholders to petition the agency to change that period. As agencies move forward with developing and implementing these polices, there will be ample opportunity for further public input to ensure they are doing the best possible job of reconciling all of the relevant interests.

In addition to addressing the issue of public access to scientific publications, the memorandum requires that agencies start to address the need to improve upon the management and sharing of scientific data produced with Federal funding. Strengthening these policies will promote entrepreneurship and jobs growth in addition to driving scientific progress. Access to pre-existing data sets can accelerate growth by allowing companies to focus resources and efforts on understanding and fully exploiting discoveries instead of repeating basic, pre-competitive work already documented elsewhere. For example, open weather data underpins the forecasting industry and provides great public benefits, and making human genome sequences publically available has spawned many biomedical innovations—not to mention many companies generating billions of dollars in revenues and the jobs that go with them. Going forward, wider availability of scientific data will create innovative economic markets for services related to data curation, preservation, analysis, and visualization, among others.

So thank you again for your petition. I hope you will agree that the Administration has done its homework and responded substantively to your request.

Dr. John Holdren is Assistant to the President for Science and Technology and Director of the White House Office of Science and Technology Policy

Tell us what you think about this response and We the People.

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

February 22, 2013

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren
Director

SUBJECT: Increasing Access to the Results of Federally Funded Scientific Research

1. Policy Principles

The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data.

Scientific research supported by the Federal Government catalyzes innovative breakthroughs that drive our economy. The results of that research become the grist for new insights and are assets for progress in areas such as health, energy, the environment, agriculture, and national security.

Access to digital data sets resulting from federally funded research allows companies to focus resources and efforts on understanding and exploiting discoveries. For example, open weather data underpins the forecasting industry, and making genome sequences publicly available has spawned many biotechnology innovations. In addition, wider availability of peer-reviewed publications and scientific data in digital formats will create innovative economic markets for services related to curation, preservation, analysis, and visualization. Policies that mobilize these publications and data for re-use through preservation and broader public access also maximize the impact and accountability of the Federal research investment. These policies will accelerate scientific breakthroughs and innovation, promote entrepreneurship, and enhance economic growth and job creation.

The Administration also recognizes that publishers provide valuable services, including the coordination of peer review, that are essential for ensuring the high quality and integrity of many scholarly publications. It is critical that these services continue to be made available. It is also important that Federal policy not adversely affect opportunities for researchers who are not funded by the Federal Government to disseminate any analysis or results of their research.

To achieve the Administration's commitment to increase access to federally funded published research and digital scientific data, Federal agencies investing in research and development must have clear and coordinated policies for increasing such access.

2. Agency Public Access Plan

The Office of Science and Technology Policy (OSTP) hereby directs each Federal agency with over $100 million in annual conduct of research and development expenditures to develop a plan to support increased public access to the results of research funded by the Federal Government. This includes any results published in peer-reviewed scholarly publications that are based on research that directly arises from Federal funds, as defined in relevant OMB circulars (e.g., A-21 and A-11). It is preferred that agencies work together, where appropriate, to develop these plans.

Each agency plan must be consistent with the objectives set out in this memorandum. These objectives were developed with input from the National Science and Technology Council and public consultation in compliance with the America COMPETES Reauthorization Act of 2010 (P.L. 111-358).

Further, each agency plan for both scientific publications and digital scientific data must contain the following elements:

a) a strategy for leveraging existing archives, where appropriate, and fostering public-private partnerships with scientific journals relevant to the agency's research;

b) a strategy for improving the public's ability to locate and access digital data resulting from federally funded scientific research;

c) an approach for optimizing search, archival, and dissemination features that encourages innovation in accessibility and interoperability, while ensuring long-term stewardship of the results of federally funded research;

d) a plan for notifying awardees and other federally funded scientific researchers of their obligations (e.g., through guidance, conditions of awards, and/or regulatory changes);

e) an agency strategy for measuring and, as necessary, enforcing compliance with its plan;

f) identification of resources within the existing agency budget to implement the plan;

g) a timeline for implementation; and

h) identification of any special circumstances that prevent the agency from meeting any of the objectives set out in this memorandum, in whole or in part.

Each agency shall submit its draft plan to OSTP within six months of publication of this memorandum. OSTP, in coordination with the Office of Management and Budget (OMB), will review the draft agency plans and provide guidance to facilitate the development of final plans that are consistent with the objectives of this memorandum and, where possible, compatible with the plans of other Federal agencies subject to this memorandum. In devising its final plan, each

agency should use a transparent process for soliciting views from stakeholders, including federally funded researchers, universities, libraries, publishers, users of federally funded research results, and civil society groups, and take such views into account.

3. Objectives for Public Access to Scientific Publications

To the extent feasible and consistent with law; agency mission; resource constraints; U.S. national, homeland, and economic security; and the objectives listed below, the results of unclassified research that are published in peer-reviewed publications directly arising from Federal funding should be stored for long-term preservation and publicly accessible to search, retrieve, and analyze in ways that maximize the impact and accountability of the Federal research investment.

In developing their public access plans, agencies shall seek to put in place policies that enhance innovation and competitiveness by maximizing the potential to create new business opportunities and are otherwise consistent with the principles articulated in section 1.

Agency plans must also describe, to the extent feasible, procedures the agency will take to help prevent the unauthorized mass redistribution of scholarly publications.

Further, each agency plan shall:

a) Ensure that the public can read, download, and analyze in digital form final peer-reviewed manuscripts or final published documents within a timeframe that is appropriate for each type of research conducted or sponsored by the agency. Specifically, each agency:

    i) shall use a twelve-month post-publication embargo period as a guideline for making research papers publicly available; however, an agency may tailor its plan as necessary to address the objectives articulated in this memorandum, as well as the challenges and public interests that are unique to each field and mission combination, and

    ii) shall also provide a mechanism for stakeholders to petition for changing the embargo period for a specific field by presenting evidence demonstrating that the plan would be inconsistent with the objectives articulated in this memorandum;

b) Facilitate easy public search, analysis of, and access to peer-reviewed scholarly publications directly arising from research funded by the Federal Government;

c) Ensure full public access to publications' metadata without charge upon first publication in a data format that ensures interoperability with current and future search technology. Where possible, the metadata should provide a link to the location where the full text and associated supplemental materials will be made available after the embargo period;

d) Encourage public-private collaboration to:

   i) maximize the potential for interoperability between public and private platforms and creative reuse to enhance value to all stakeholders,

   ii) avoid unnecessary duplication of existing mechanisms,

   iii) maximize the impact of the Federal research investment, and

   iv) otherwise assist with implementation of the agency plan;

e) Ensure that attribution to authors, journals, and original publishers is maintained; and

f) Ensure that publications and metadata are stored in an archival solution that:

   i) provides for long-term preservation and access to the content without charge,

   ii) uses standards, widely available and, to the extent possible, nonproprietary archival formats for text and associated content (e.g., images, video, supporting data),

   iii) provides access for persons with disabilities consistent with Section 508 of the Rehabilitation Act of 1973,[1] and

   iv) enables integration and interoperability with other Federal public access archival solutions and other appropriate archives.

Repositories could be maintained by the Federal agency funding the research, through an arrangement with other Federal agencies, or through other parties working in partnership with the agency including, but not limited to, scholarly and professional associations, publishers and libraries.

4. Objectives for Public Access to Scientific Data in Digital Formats

To the extent feasible and consistent with applicable law and policy[2]; agency mission; resource constraints; U.S. national, homeland, and economic security; and the objectives listed below, digitally formatted scientific data resulting from unclassified research supported wholly or in part

[1] Section 508 Of The Rehabilitation Act, as amended, available at: https://www.section508.gov/index.cfm?fuseAction=1998Amend

[2] These policies include, but are not limited to OMB Circular A-130, Management of Federal Information Resources, available at: http://www.whitehouse.gov/omb/circulars_a130_a130trans4

by Federal funding should be stored and publicly accessible to search, retrieve, and analyze. For purposes of this memorandum, data is defined, consistent with OMB circular A-110, as the digital recorded factual material commonly accepted in the scientific community as necessary to validate research findings including data sets used to support scholarly publications, but does not include laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects, such as laboratory specimens. Each agency's public access plan shall:

a) Maximize access, by the general public and without charge, to digitally formatted scientific data created with Federal funds, while:

   i) protecting confidentiality and personal privacy,

   ii) recognizing proprietary interests, business confidential information, and intellectual property rights and avoiding significant negative impact on intellectual property rights, innovation, and U.S. competitiveness, and

   iii) preserving the balance between the relative value of long-term preservation and access and the associated cost and administrative burden;

b) Ensure that all extramural researchers receiving Federal grants and contracts for scientific research and intramural researchers develop data management plans, as appropriate, describing how they will provide for long-term preservation of, and access to, scientific data in digital formats resulting from federally funded research, or explaining why long-term preservation and access cannot be justified;

c) Allow the inclusion of appropriate costs for data management and access in proposals for Federal funding for scientific research;

d) Ensure appropriate evaluation of the merits of submitted data management plans;

e) Include mechanisms to ensure that intramural and extramural researchers comply with data management plans and policies;

f) Promote the deposit of data in publicly accessible databases, where appropriate and available;

g) Encourage cooperation with the private sector to improve data access and compatibility, including through the formation of public-private partnerships with foundations and other research funding organizations;

h) Develop approaches for identifying and providing appropriate attribution to scientific data sets that are made available under the plan;

i) In coordination with other agencies and the private sector, support training, education, and workforce development related to scientific data management, analysis, storage, preservation, and stewardship; and

j) Provide for the assessment of long-term needs for the preservation of scientific data in fields that the agency supports and outline options for developing and sustaining repositories for scientific data in digital formats, taking into account the efforts of public and private sector entities.

5. Implementation of Public Access Plans

Some Federal agencies already have policies that partially meet the requirements of this memo. Those agencies should adapt those policies, as necessary, to fully meet the requirements. Once finalized, each agency should post its public access plan on its Open Government website.

The agency plan shall not apply to manuscripts submitted for publication prior to the plan's effective date or to digital data generated prior to the plan's effective date. The effective dates can be no sooner than the publication date of the agency's final plan.

OSTP will oversee implementation through regular meetings with agencies. Each agency shall provide updates on implementation to the Directors of OSTP and OMB twice yearly; these updates shall be submitted by January 1 and July 1 of each year for two years after the effective date of the agency's final plan. An agency may amend its public access plan consistent with these objectives, in consultation with OSTP and OMB.

6. General Provisions

Nothing in this memorandum shall be construed to impair or otherwise affect authority granted by law to an executive department, agency, or the head thereof; or functions of the Director of OMB relating to budgetary, administrative, or legislative proposals.

Consistent with the America COMPETES Reauthorization Act of 2010, nothing in this memorandum, or the agency plans developed pursuant to it, shall be construed to authorize or require agencies to undermine any right under the provisions of title 17 or 35, United States Code, or to violate the international obligations of the United States. This memorandum is not intended to, and does not, create any right or benefit, substantive or procedural, enforceable at law or in equity, by any party against the United States; its departments, agencies; or entities, its officers, employees, or agents; or any other person.

APPENDIX D. OPEN DATA POLICY MEMORANDUM

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF MANAGEMENT AND BUDGET
WASHINGTON, D.C. 20503

THE DIRECTOR

May 9, 2013

M-13-13

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Sylvia M. Burwell
Director

Steven VanRoekel
Federal Chief Information Officer

Todd Park
U.S. Chief Technology Officer

Dominic J. Mancini
Acting Administrator, Office of Information and Regulatory Affairs

SUBJECT: Open Data Policy-Managing Information as an Asset

Information is a valuable national resource and a strategic asset to the Federal Government, its partners, and the public. In order to ensure that the Federal Government is taking full advantage of its information resources, executive departments and agencies (hereafter referred to as "agencies") must manage inf01mation as an asset throughout its life cycle to promote openness and interoperability, and properly safeguard systems and information. Managing government information as an asset will increase operational efficiencies, reduce costs, improve services, support mission needs, safeguard personal information, and increase public access to valuable government information.

Making information resources accessible, discoverable, and usable by the public can help fuel entrepreneurship, innovation, and scientific discovery-all of which improve Americans' lives and contribute significantly to job creation. For example, decades ago, the Federal Government made both weather data and the Global Positioning System (GPS) freely available to anyone. Since then, American entrepreneurs and innovators have used these resources to create navigation systems, weather newscasts and warning systems, location-based applications, precision farming tools, and much more.

Pursuant to Executive Order of May 9, 2013, *Making Open and Machine Readable the New Default for Government Information,* this Memorandum establishes a framework to help institutionalize the principles of effective information management at each stage of the information's life cycle to promote interoperability and openness. Whether or not particular information can be made public, agencies can apply this framework to all information resources to promote efficiency and produce value.

Specifically, this Memorandum requires agencies to collect or create information in a way that supports downstream information processing and dissemination activities. This includes using machine-readable and open formats, data standards, and common core and extensible metadata for all new

information creation and collection efforts. It also includes agencies ensuring information stewardship through the use of open licenses and review of information for privacy, confidentiality, security, or other restrictions to release. Additionally, it involves agencies building or modernizing information systems in a way that maximizes interoperability and information accessibility, maintains internal and external data asset inventories, enhances information safeguards, and clarifies information management responsibilities.

The Federal Government has already made significant progress in improving its management of information resources to increase interoperability and openness. The President's Memorandum on *Transparency and Open Government*[1] instructed agencies to take specific actions to implement the principles of transparency, participation, and collaboration, and the Office of Management and Budget's (OMB) *Open Government Directive*[2] required agencies to expand access to information by making it available online in open formats. OMB has also developed policies to help agencies incorporate sound information practices, including OMB Circular A-130[3] and OMB Memorandum M-06-02.[4] In addition, the Federal Government launched Data.gov, an online platform designed to increase access to Federal datasets. The publication of thousands of data assets through Data.gov has enabled the development of numerous products and services that benefit the public.

To help build on these efforts, the President issued a Memorandum on May 23, 2012 entitled *Building a 21st Century Digital Government*[5] that charged the Federal Chief Information Officer (CIO) with developing and implementing a comprehensive government-wide strategy to deliver better digital services to the American people. The resulting *Digital Government Strategy*[6] outlined an information-centric approach to transform how the Federal Government builds and delivers digital services, and required OMB to develop guidance to increase the interoperability and openness of government information.

This Memorandum is designed to be consistent with existing requirements in the Paperwork Reduction Act,[7] the E-Government Act of 2002,[8] the Privacy Act of 1974,[9] the Federal Information Security Management Act of 2002 (FISMA),[10] the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA),[11] the Freedom of Information Act,[12] the Information Quality Act,[13] the

---

[1] President Barack Obama, Memorandum on Transparency and Open Government (Jan. 21, 2009), *available at* http://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment.

[2] OMB Memorandum M-10-06, *Open Government Directive* (Dec. 8, 2009), *available at* http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-06.pdf

[3] OMB Circular A-130, *available at* http://www.whitehouse.gov/omb/Circulars_a130_a130trans4

[4] OMB Memorandum M-06-02, *Improving Public Access to and Dissemination of Government Information and Using the Federal Enterprise Architecture Data Reference Model* (Dec. 16, 2005), *available at* http://www.whitehouse.gov/sites/default/files/omb/memoranda/fy2006/m06-02.pdf

[5] President Barack Obama, Memorandum on Building a 21st Century Digital Government (May 23, 2012), *available at* http://www.whitehouse.gov/sites/default/files/uploads/2012digital_mem_rel.pdf

[6] Office of Management and Budget, *Digital Government: Building a 21st Century Platform to Better Serve the American People* (May 23, 2012), *available at* http://www.whitehouse.gov/sites/default/files/omb/egov/digital-government/digital-government-strategy.pdf

[7] 44 U.S.C. § 3501 *et seq.*

[8] Pub. L. No. 107-347, 116 Stat. 2899 (2002) (codified as 44 U.S.C. § 3501 note).

[9] 5 U.SC. § 552a.

[10] 44 U.S.C. § 3541, *et seq.*

[11] Section 503(a), Pub. L. No. 107-347, 116 Stat. 2899 (2002) (codified as 44 U.S.C. § 3501 note); *see also* Implementation Guidance for Title V of the E-Government Act, Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA), *available at* http://www.whitehouse.gov/sites/default/files/omb/assets/omb/fedreg/2007/061507_cipsea_guidance.pdf

[12] 5 USC 552(a)(2).

Federal Records Act, [14] and existing OMB and Office of Science and Technology Policy (OSTP) guidance.

If agencies have any questions regarding this Memorandum, please direct them to OMB at datause@omb.eop.gov.

Attachment

[13] Pub. L. No. 106-554 (2000) (codified at 44 U.S.C. § 3504(d)(l) a nd 35 1 6). *See also* OMB Memorandum M-12-18, *Managing Government Records Directive* (Aug. 24, 20 1 2), *available at* http://www.whitehouse.gov/sites/default/files/omb/memoranda/2012/m-12-18.pdf.

[14] 44 U.S.C. Chapters 21, 22, 29, 31, and 33. *See also* 36 CFR Subchapter B - Records Management.

Attachment

This attachment provides definitions and implementation guidance for M-13-13, *Open Data Policy-Managing Information as an Asset.*

I. Definitions:

Data: For the purposes of this Memorandum, the term "data" refers to all structured information, unless otherwise noted. [15]

Dataset: For the purposes of this Memorandum, the term "dataset" refers to a collection of data presented in tabular or non-tabular form.

Fair Information Practice Principles: The term "Fair Information Practice Principles" refers to the eight widely accepted principles for identifying and mitigating privacy impacts in information systems, program s and processes, delineated in the National Strategy for Trusted Identities in Cyberspace.[16]

Government information: As defined in OMB Circular A-130, "government information" means information created , collected, processed , disseminated, or disposed of, by or for the Federal Government.

Information: As defined in OMB Circular A-130, the term "information" mea ns any communication or re presentation of knowledge such as facts, data, or opinions in any medium or form, including textual, numerical, graphic, cartographic, narrative, or audiovisual forms.

Information life cycle: As defined in OMB Circular A-130, the term "information life cycle" means the stages through which information passes, typically characterized as creation or collection, processing, dissemination , use, storage, and disposition.

Personally identifiable information: As defined in OMB Memorandum M-10-23, [17] "personally identifiable information" (PII) refers to information that can be used to distinguish or trace an individual 's identity, either alone or when combined with other personal or identifying information that is linked or linkable to a specific individual. The definition of PII is not anchored to any single category of information or technology. Rather, it requires a case-by-case assessment of the specific risk that an individual can be identified. In performing this assessment, it is important for an agency to recognize that non-PII can become PII whenever additional information is made publicly available (in any medium and from any source) that, when combined with other available information , could be used to identify an individual.

Mosaic effect: The mosaic effect occurs when the information in an individual dataset, in isolation , may not pose a risk of identifying an individual (or threaten in g some other important interest such as security), but when combined with other available information, could pose such risk. Before disclosing potential PII or other potentially sensitive information, agencies must consider other publicly available data- in

---

[15] *Structured* information is to be contrasted with *unstructured* information (commonly referred to as "content") such as press releases and fact sheets. As described in the *Digital Government Strategy,* content may be converted to a structured format and treated as data. For example, a web-based fact sheet may be broken into the following component data pieces: the title, body text, images, and related links.

[16] The White House, *National Strategy for Trusted Identities in Cyberspace* (Apr il 20 II), *available at* http:// www.whitehouse.gov/sites/default/files/rss_viewer/NSTICstrategy_041511.pdf

[17] OMB Memorandum M-1 0-23, *Guidance for Agency Use of Third-Party Websites and Applications* (June 25, 20 I 0), *available at* http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-23.pdf

any medium and from any source- to determine whether some combination of existing data and the data intended to be publicly released could allow for the identification of an individual or pose another security concern .

Open data: For the purposes of this Memorandum, the term "open data" refers to publicly available data structured in a way that enables the data to be fully discoverable and usable by end users. In general, open data will be consistent with the following principles:

- *Public.* Consistent with OMB's *Open Government Directive,* agencies must adopt a presumption in favor of openness to the extent permitted by law and subject to privacy, confidentiality, security, or other valid restrictions.

- *Accessible.* Open data are made available in convenient, modifiable, and open formats that can be retrieved, downloaded, indexed, and searched. Formats should be machine-readable (i.e., data are reasonably structured to allow automated processing). Open data structures do not discriminate against any person or group of persons and should be made available to the widest range of users for the widest range of purposes, often by providing the data in multiple formats for consumption. To the extent permitted by law, these formats should be non-proprietary, publicly available, and no restrictions should be placed upon their use.

- *Described.* Open data are described fully so that consumers of the data have sufficient information to understand their strengths, weaknesses, analytical limitations, security requirements, as well as how to process them. This involves the use of robust, granular metadata (i.e., fields or elements that describe data), thorough documentation of data elements, data dictionaries, and, if applicable, additional descriptions of the purpose of the collection, the population of interest, the characteristics of the sample, and the method of data collection.

- *Reusable.* Open data are made available under an open license that places no restrictions on their use.

- *Complete.* Open data are published in primary forms (i.e., as collected at the source), with the finest possible level of granularity that is practicable and permitted by law and other requirements. Derived or aggregate open data should also be published but must reference the primary data.

- *Timely.* Open data are made available as quickly as necessary to preserve the value of the data. Frequency of release should account for key audiences and downstream needs.

- *Managed Post-Release.* A point of contact must be designated to assist with data use and to respond to complaints about adherence to these open data requirements.

Project Open Data: "Project Open Data," a new OMB and OSTP resource, is an online repository of tools, best practices, and schema to help agencies adopt the framework presented in this guidance. Project Open Data can be accessed at http://project-open-data.github.io. [18] Project Open Data will evolve over time as a community resource to facilitate adoption of open data practices. The repository includes definitions, code, checklists, case studies, and more, and enables collaboration across the Federal Government, in partnership with public developers, as applicable. Agencies can visit Project Open Data for a more comprehensive glossary of terms related to open data.

[18] Links to the best practices developed in Project Open Data referenced in this memorandum can be found through the directory on this main page.

II. **Scope:**

The requirements in part III, sections 1 and 2 of this Memorandum apply to all new information collection , creation, and system development efforts as well as major modernization projects that update or re-design existing information system s. National Security Systems, as defined in 40 U.S. C. 11103, are exempt from the requirements of this policy . The requirements in part III, section 3 apply to management of all datasets used in an agency's information system s. Agencies are also encouraged to improve the discoverability and usability of existing datasets by making them "open" using the methods outlined in this Memorandum, prioritizing those that have already been released to the public or otherwise deemed high-value or high-demand through engagement with customers (see part III, sect ion 3.c). Agencies should exercise judgment before publicly distributing data residing in an existing system by weighing the value of openness against the cost of making those data public.

III. **Policy Requirements**:

Agencies management of information resources must begin at the earliest stages of the planning process, well before information is collected or created. Early strategic planning will allow the Federal Government to design systems and develop processes that unlock the full value of the information, and provide a foundation from which agencies can continue to manage information throughout its life cycle.

Agencies shall take the follow in g actions to improve the management of information resources throughout the information's life cycle and reinforce the government 's presumption in favor of openness:

1. **Collect or create information in a way that supports downstream information processing and dissemination activities**- Consistent with OMB Circular A-130, agencies must consider, at each stage of the information life cycle, the effects of decisions and actions on other stages of the life cycle. Accordingly, to the extent permitted by law, agencies must design new information collection and creation efforts so that the information collected or created supports downstream interoperability between information system s and dissemination of information to the public, as appropriate, without the need for costly retrofitting. This includes consideration and consultation of key target audiences for the information when determining format, frequency of update, and other information management decisions. Specifically, agencies must incorporate the following requirements into future information collect ion and creation efforts:

    a. **Use machine-readable and open formats** [19] - Agencies must use machine-readable and open formats for information as it is collected or created. While information should be collected electronically by default, machine-readable and open formats must be used in conjunction with both electronic and telephone or paper-based information collect ion efforts. Additionally, in consultation with the best practices found in Project Open Data and to the extent permitted by l aw, agencies should prioritize the u se of open formats that are non-proprietary, publicly available, and that place no restrictions upon their u se.

[19] The requirements of this subsection build upon existing requirements in OMB Statistical Policy Directives No. l and No.2, *available at* http://www.whitehouse.gov/sitcs/dcfault/files/omb/assets/omb/inforeg/directivel.pdf and http://www.whitehouse.gov/sites/default/files/omb/assets/omb/inforeg/directive2.pdf.

b. **Use data standards**- Consistent with existing policies relating to Federal agencies' use of standards[20] for information as it is collected or created , agencies must use standards in order to promote data interoperability and openness.

c. **Ensure information stewardship through the use of open licenses** - Agencies must apply open licenses, in consultation with the best practices found in Project Open Data, to information as it is collected or created so that if data are made public there are no restrictions on copying, publishing, distributing, transmitting, adapting, or otherwise using the information for non-commercial or for commercial purposes.[21] When information is acquired or accessed by an agency through performance of a contract, appropriate existing clauses[22] shall be utilized to meet these objectives while recognizing that contractors may have proprietary interests in such information , and that protection of such information may be necessary to encourage qualified contractors to participate in and apply innovative concepts to government programs.

d. **Use common core and extensible metadata**- Agencies must describe information using common core metadata, in consultation with the best practices found in Project Open Data, as it is collected and created. Metadata should also include information about origin, linked data, geographic location, time series continuations, data quality, and other relevant indices that reveal relationships between datasets and allow the public to determine the fitness of the data source. Agencies may expand upon the basic common metadata based on standards, specifications, or formats developed within different communities (e.g., financial, health, geospatial, law enforcement). Groups that develop and promulgate these metadata specifications must review them for compliance with the common core metadata standard , specifications, and formats.

2. **Build information systems to support interoperability and information accessibility**- Through their acquisition and technology management processes, agencies must build or modernize information systems in a way that maximizes interoperability and information accessibility, to the extent practicable and permitted by law. To this end, agencies should leverage existing Federal IT guidance, such as the *Common Approach to Federal Enterprise Architecture,*[23] when designing information systems. Agencies must exercise forethought when architecting, building, or substantially modifying an information system to facilitate public distribution, where appropriate. In addition, the agency 's CIO must validate that the following minimum requirements have been incorporated into acquisition planning documents and technical design for all new information systems and those preparing for modernization, as appropriate:

a. The system design must be scalable, flexible, and facilitate extraction of data in multiple formats and for a range of uses as internal and external needs change, including potential uses not accounted for in the original design. In general , this will involve the use of standards and specifications in the system design that promote industry best practices for information

---

[20] *See* OMB Circular A-1 I 9, *available at* http://www.whitehouse.gov/omb/circulars_a119, and OMB Memorandum M-12-08, *Principles for Federal Engagement in Standards Activities to Address National Priorities* (Jan 27, 2012), *available at* http://www.whitehouse.gov/sites/default/files/omb/memoranda/2012/m-12-08.pdf

[21] If a data user augments or alters original information that is attributed to the Federal Government, the user is responsible for making clear the source and nature of that augmentation .

[22] *See* Federal Acquisition Regulation (FAR) Subpart 27.4-Rights in Data and Copyrights, *available at* https://acquisition.gov/far/current/html/Subpart%2027_4. html

[23] Office of Management and Budget, Common Approach to Federal Enterprise Architecture, *available at* http://www.whitehouse.gov/sites/default/files/omb/assets/egov_docs/common_*approach*_to_federal_ea.pdf

sharing, and separation of data from the application layer to maximize data reuse opportunities and incorporation of future application or technology capabilities, in consultation with the best practices found in Project Open Data;

b. All data outputs associated with the system must meet t he requirements described in part III, sections l .a-e of this Memorandum and be accounted for in the data inventory described in part III section 3.a; and

c. Data schema and dictionaries have been documented and shared with internal partners and the public, as applicable.

3. **Strengthen data management and release practices-** To ensure that agency data assets are managed and maintained throughout their life cycle, agencies must adopt effective data asset portfolio management approaches. Within six (6) months of the date of this Memorandum, agencies and interagency groups must review and, where appropriate, revise existing policies and procedures to strengthen their data management and release practices to ensure consistency with the requirements in this Memorandum, and take the following actions:

   a. **Create and maintain an enterprise data inventory-** Agencies must update their inventory of agency information resources (as required by OMB Circular A-1 30)[24] to include an enterprise data inventory, if it does not al ready exist, that accounts for datasets used in the agency's information systems. The inventory will be built out over time, with the ultimate goal of including all agency datasets, to the extent practicable. The inventory will indicate, as appropriate, if the agency has determined that the individual datasets may be made publicly available (i.e., release is permitted by law, subject to all privacy, confidentiality, security, and other valid requirements) and whether they are current l y available to the public. The Senior Agency Official for Records Management should be consulted on integration with the records management process. Agencies should use the Data Reference Model from the Federal Enterprise Architecture[25] to help create and maintain their inventory. Agencies must describe datasets within the inventory u sing the common core and extensible metadata (see part III, section l.e).

   b. **Create and maintain a public data listing-** Any datasets in the agency's enterprise data inventory that can be made publicly available must be listed at www.[agency].gov/data in a human- and machine-readable format that enables automatic aggregation by Data.gov and other services (known as "harvestable files"), to the extent practicable. This should include datasets that can be made publicly available but have not yet been released. This public data li sting should al so include, to the extent permitted by law and existing terms and conditions, datasets that were produced through agency-funded grants, contracts, and cooperative agreements (excluding any data submitted primarily for the purpose of contract monitoring and administration), and , where feasible, be accompanied by standard citation information , preferably in the form of a persistent identifier. The public data listing will be built out over time, with the ultimate goal of including all agency datasets that can be made publicly available. See Project Open Data for best practices, tools, and schema to implement the public data li sting and harvestable files.

---

[24] *See* OMS Circular A-130, secti on 8(b)(2)(a).

[25] Office of Management and Bud get, Federal Enterprise Architecture (FEA) Reference Models, *available at* http://www.whitehouse.gov/omb/e-gov/fea

c. **Create a process to engage with customers to help facilitate and prioritize data release-** Agencies must create a process to engage with customers, through their www.[agency].gov/data pages and other necessary means, to solicit help in prioritizing the release of datasets and determining the most usable and appropriate formats for release?[6]
Agencies should make data available in multiple formats according to their customer needs. For example, high-volume datasets of interest to developers should be released using bulk downloads as well as Application Programming Interfaces (APIs). In addition, customer engagement efforts should help agencies prioritize efforts to improve the discoverability and usability of datasets that have already been released to the public but are not yet fully "open" (e.g., they are only available in closed, inaccessible formats). See Project Open Data for best practices and tools that can be used to implement customer engagement efforts.

d. **Clarify roles and responsibilities for promoting efficient and effective data release practices-** Agencies must ensure that roles and responsibilities are clearly designated for the promotion of efficient and effective data release practices across the agency, and that proper authorities have been granted to execute on related responsibilities, including:

1. Communicating the strategic value of open data to internal stakeholders and the public;
11. Ensuring that data released to the public are open (as defined in part I), as appropriate, and a point of contact is designated to assist open data use and to respond to complaints about adherence to open data requirements;
111. Engaging entrepreneurs and innovators in the private and nonprofit sectors to encourage and facilitate the use of agency data to build applications and services;
IV. Working with agency components to scale best practices from bureaus and offices that excel in open data practices across the enterprise;
v. Working with the agency 's Senior Agency Official for Privacy (SAOP) or other relevant officials to ensure that privacy and confidentiality are fully protected; and
vt. Working with the Chief Information Security Officer (CISO) and mission owners to assess overall organizational risk, based on the impact of releasing potentially sensitive data, and make a risk-based determination.

4. **Strengthen measures to ensure that privacy and confidentiality are fully protected and that data are properly secured-** Agencies must incorporate privacy analyses into each stage of the information 's life cycle. In particular, agencies must review the information collected or created for valid restrictions to release to determine whether it can be made publicly available, consistent with the *Open Government Directive's* presumption in favor of openness, and to the extent permitted by law and subject to privacy, confidentiality pledge, security, trade secret, contractual , or other valid restrictions to release. If the agency determines that information should not be made publicly available on one of these grounds, the agency must document this determination in consultation with its Office of General Counsel or
equivalent.

As agencies consider whether or not information may be disclosed, they must also account for the "mosaic effect" of data aggregation. Agencies should note that the mosaic effect demands a risk-based

[26] OMB Statistical Policy Directives 3 and 4 describe the schedule and manner in which data produced by the principal statistical agencies will be released. Statistical Policy Directive No.4: Release and Dissemination of Statistical Products Produced by Federal Statistical Agencies, *available at* http://www.whitehouse.gov/sites/default/files/omb/assets/omb/fedreg/2008/030708_directive-4.pdf; Statistical Policy Directive 3: Compilation, Release, and Evaluation of Principal Federal Economic Indicators, *available at* http://www.whitehouse.gov/sites/defaultlfiles/omb/assets/omb/inforeg/statpolicy/dir_3_fr_09251985.pdf

analysis, often utilizing statistical methods whose parameters can change over time, depending on the nature of the information, the availability of other information, and the technology in place that could facilitate the process of identification. Because of the Complexity of this analysis and the scope of data involved, agencies may choose to take advantage of entities in the Executive Branch that may have relevant expertise, including the staff of Data.gov. Ultimately, it is the responsibility of each agency to perform the necessary analysis and comply with all applicable laws, regulations, and policies. In some cases, this assessment may affect the amount, type, form, and detail of data released by agencies.

As OMB has noted, "The individual 's right to privacy must be protected in Federal Government information activities involving personal information."[27] As agencies consider security-related restrictions to release, they should focus on information confidentiality, integrity, and availability as part of the agency's overall risk management framework. They are required to incorporate the National Institute of Standards and Technology (NIST) Federal Information Processing Standard (FIPS) Publication 199 "Standards for Security Categorization of Federal Information and Information Systems," which includes guidance and definitions for confidentiality, integrity, and availability.[28] Agencies should also consult with the Controlled Unclassified Information (CUI) program to ensure compliance with CUI requirements,[29] the National Strategy for Information Sharing and Safeguarding,[30] and the best practices found in Project Open Data. In addition to complying with the Privacy Act of 1974, the E-Government Act of2002, FISMA, and CIPSEA, agencies should implement information policies based upon Fair Information Practice Principles and NIST guidance on Security and Privacy Controls for Federal Information Systems and Organizations.[31] For example, agencies must:

a. Collect or create only that information necessary for the proper performance of agency functions and which has practical utility;[32]
b. Limit the collection or creation of information which identifies individuals to that which is legally authorized and necessary for the proper performance of agency functions;[33]
c. Limit the sharing of information that identifies individuals or contains proprietary information to that which is legally authorized, and impose appropriate conditions on use where a continuing obligation to ensure the confidentiality of the information exists;[34]
d. Ensure that information is protected commensurate with the risk and magnitude of the harm that would result from the loss, misuse, or unauthorized access to or modification of such information;[35] and
e. Take into account other publicly available information when determining whether particular information should be considered PII (as defined in part I of this Memorandum).

[27] *See* OMB Circular A-130, *available at* http:// www.whitehouse.gov/omb/Circulars_al30_a130trans4/

[28] NIST FIPS Publication 199 "Standards for Security Categorization of Federal Information and Information Systems", *available at* http://csrc.nist.gov/publications/fips/fips199/FIPS-PUB-199-final.pdf

[29] Executive Order 13556, Controlled Unclassified Information , *available at* http://www.whitehouse.gov/the-press-office/2010/11/04/executive-order-13556-controlled-unclassifled-information.

[30] The White House, *National Strategy for Information Sharing and Safeguarding* (December 2011), *available at* http://www.whitehouse.gov/the-press-officc/2012/12/19/ national-strategy-information-sharing-and-safeguarding

[31] *See* NIST Special Publication 800-53 "Security and Privacy Controls for Federal Information Systems and Organizations", *available at* http://csrc.nist.gov/publications/drafts/800-53-rcv4/sp800-53-rev4 -ipd.pdf

[31] *See* OMB Circ ul a r A-130, secti on 8(a)(2).

[33] *See* OMB Circular A -130, section 8(a)(9)(b).

[34] *See* OMB Circul a r A -130, secti on 8(a)(9)(c).

[35] *See* OMB Circul ar A-130, secti on 8(a)(9)(a).

5. **Incorporate new interoperability and openness requirements into core agency processes-** Consistent with 44 U.S.C. 3506 (b)(2), agencies must develop and maintain an Information Resource Management (IRM) Strategic Plan. IRM Strategic Plans should align with the agency 's Strategic Plan (as required by OMB Circular A-11),[36] support the attainment of agency priority goals as mandated by the Government Performance and Results Modernization Act of 2010,[37] provide a description of how IRM activities help accomplish agency missions, and ensure that IRM decisions are integrated with organizational planning, bud get, procurement, financial management, human resources management, and program decisions. As part of the annual PortfolioStat process,[38] agencies must update their IRM Strategic Plans to describe how they are meeting new and existing information life cycle management requirements. Specifically, agencies must describe how they have institutionalized and operationalized the interoperability and openness requirements in this Memorandum into their core processes across all applicable agency programs and stakeholders.

IV. **Implementation**:

As agencies take steps to meet the requirements in this Memorandum, it is important to work strategically and prioritize those elements that can be addressed immediately, support mission-critical objectives, and result in more efficient use of tax payer dollars.

Agencies should consider the following as they implement the requirements of this Memorandum:

1. **Roles and Responsibilities-** The Clinger-Cohen Act of 1996 assigns agency CIOs statutory responsibility for promoting the effective and efficient design and operation of all major IRM processes within their agency. Accordingly, agency head s must ensure that CIOs are positioned with the responsibility and authority to implement the requirements of this Memorandum in coordination with the agency's Chief Acquisition Officer, Chief Financial Officer, Chief Technology Officer, Senior Agency Official for Geospatial Information , Senior Agency Official for Privacy (SAOP), Chief Information Security Officer (CISO), Senior Agency Official for Records Management, and Chief Freedom of Information Act (FOIA) Officer. The CIO should also work with the agency's public affairs staff, open government staff, web manager or digital strategist, program owners and other leadership, as applicable.

   A key component of agencies' management of information resources involves working closely with the agency's SAOP and other relevant officials to ensure that each stage of the planning process includes a full analysis of privacy, confidentiality, and security issues. Agency heads must also ensure that privacy and security officials are positioned with the authority to identify information that may require additional protection and agency activities that may require additional safeguards. Consistent with OMB Memorandum M-05-08,[39] each agency 's SAOP must take on a central planning and policy-making role in all agency information management activities, beginning at the earliest stages of planning and continuing throughout the life cycle of the information. In addition, if an agency's SAOP is not positioned within the office of the CIO, the agency should designate an official within the office of the CIO to serve as a liaison to help coordinate with the agency 's privacy office.

---

[36] OMB Circular A-ll , *available at* http://www.w hitehouse.gov/omb/circulars_all_current_year_all_toe

[37] Pub. L. No. 111-352 (2011) (codified as 3 1 USC§ 11 20 note).

[38] In March 201 2 OMB established PortfolioStat accountability sessions, engaging direct l y with agency leadership to assess the maturity and effectiveness of current IT management practices and address management opportunities and challenges. For FY 1 3 OMB PortfolioStat guidance, *see* OMB Memorandum M-13-09, *Fiscal Year 2013 PorifolioStat Guidance: Strengthening Federal IT Porifolio Management* (Mar. 27, 2013), *available at* http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-09.pdf.

[39] OMB Memorandum M -05-08, *Designation of Senior Agency Officials for Privacy* (Fe b. II , 2005), *available at* http://m.whitehouse.gov/sites/default/files/omb/assets/omb/memoranda/fy2005/m05-08.pdf

2. **Government-wide Coordination-** The Federal CIO will work with the United States Chief Technology Officer (CTO) and the Administrator of the OMB Office of Information and Regulatory Affairs (OIRA) to help improve the interoperability and openness of government information. To this end, the Federal CIO will work to establish an interagency working group supplied by the Federal CIO Council. The working group should focus on leveraging government-wide communities of practice to help with the development of tools that support information interoperability and openness through repositories such as Project Open Data. Part of this work should be to share best practices related to interoperability and openness within government (e.g., Federal, state, local, and tribal). These collaborations shall be subject to statutory limitations and conducted in a way that fully protects privacy, confidentiality, confidential business information, and intellectual property rights.

3. **Resources-** Policy implementation may require upfront investments depending on the maturity of existing information life cycle management processes at individual agencies. Agencies are encouraged to evaluate current processes and identify implementation opportunities that may result in more efficient use of taxpayer dollars. However, effective implementation should result in downstream cost savings for the enterprise through increased interoperability and accessibility of the agency's information resources. Therefore, these potential upfront investments should be considered in the context of their future benefits and be funded appropriately through the agency's capital planning and budget processes. Some of the requirements in this Memorandum may require additional tools and resources. Agencies should make progress commensurate with available tools and resources.

   In addition, tools, best practices, and schema to help agencies implement the requirements of this Memorandum can be found through the Digital Services Innovation Center and in Project Open Data.

4. **Accountability Mechanisms-** Progress on agency implementation of the actions required in this Memorandum will be primarily assessed by OMB and the public through analysis of the agency's updates to IRM plans (part III, section 5), the completeness of the enterprise data inventory (part III, section 3.a), and the data made available in the agency's public data listing (part III, section 3.b).

   Nothing in this Memorandum shall be construed to affect existing requirements for review and clearance of pre-decisional inf01mation by OMB relating to legislative, budgetary, administrative, and regulatory materials. Moreover, nothing in this Memorandum shall be construed to reduce the protection of information whose release would threaten national security, invade personal privacy, breach confidentiality or contractual terms, violate the Trade Secrets Act,[41] violate other statutory confidentiality

requirements,[41] or damage other compelling interests. This Memorandum is not intended to, and does not, create any right or benefit, substantive or procedural, enforceable at law or in equity by any party against the United States, its departments, agencies, or entities, its officers, employees, or agents, or any other person.

[40] 18 USC § 1905.

[41] *See* 13 U.S.C. §§ 8, 9 and 301(g) and 22 U.S.C. § 3104.

APPENDIX E. EXECUTIVE ORDER 13642

AUTHENTICATED U.S. GOVERNMENT INFORMATION GPO

28111

Federal Register
Vol. 78, No. 93
Tuesday, May 14, 2013

Presidential Documents

Title 3—

The President

Executive Order 13642 of May 9, 2013

Making Open and Machine Readable the New Default for Government Information

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

Section 1. *General Principles*. Openness in government strengthens our democracy, promotes the delivery of efficient and effective services to the public, and contributes to economic growth. As one vital benefit of open government, making information resources easy to find, accessible, and usable can fuel entrepreneurship, innovation, and scientific discovery that improves Americans' lives and contributes significantly to job creation.

Decades ago, the U.S. Government made both weather data and the Global Positioning System freely available. Since that time, American entrepreneurs and innovators have utilized these resources to create navigation systems, weather newscasts and warning systems, location-based applications, precision farming tools, and much more, improving Americans' lives in countless ways and leading to economic growth and job creation. In recent years, thousands of Government data resources across fields such as health and medicine, education, energy, public safety, global development, and finance have been posted in machine-readable form for free public use on Data.gov. Entrepreneurs and innovators have continued to develop a vast range of useful new products and businesses using these public information resources, creating good jobs in the process.

To promote continued job growth, Government efficiency, and the social good that can be gained from opening Government data to the public, the default state of new and modernized Government information resources shall be open and machine readable. Government information shall be managed as an asset throughout its life cycle to promote interoperability and openness, and, wherever possible and legally permissible, to ensure that data are released to the public in ways that make the data easy to find, accessible, and usable. In making this the new default state, executive departments and agencies (agencies) shall ensure that they safeguard individual privacy, confidentiality, and national security.

Sec. 2. *Open Data Policy*. (a) The Director of the Office of Management and Budget (OMB), in consultation with the Chief Information Officer (CIO), Chief Technology Officer (CTO), and Administrator of the Office of Information and Regulatory Affairs (OIRA), shall issue an Open Data Policy to advance the management of Government information as an asset, consistent with my memorandum of January 21, 2009 (Transparency and Open Government), OMB Memorandum M–10–06 (Open Government Directive), OMB and National Archives and Records Administration Memorandum M–12–18 (Managing Government Records Directive), the Office of Science and Technology Policy Memorandum of February 22, 2013 (Increasing Access to the Results of Federally Funded Scientific Research), and the CIO's strategy entitled ''Digital Government: Building a 21st Century Platform to Better Serve the American People.'' The Open Data Policy shall be updated as needed.

(b) Agencies shall implement the requirements of the Open Data Policy and shall adhere to the deadlines for specific actions specified therein. When implementing the Open Data Policy, agencies shall incorporate a full analysis of privacy, confidentiality, and security risks into each stage

of the information lifecycle to identify information that should not be released. These review processes should be overseen by the senior agency official for privacy. It is vital that agencies not release information if doing so would violate any law or policy, or jeopardize privacy, confidentiality, or national security.

Sec. 3. *Implementation of the Open Data Policy.* To facilitate effective Government-wide implementation of the Open Data Policy, I direct the following:

(a) Within 30 days of the issuance of the Open Data Policy, the CIO and CTO shall publish an open online repository of tools and best practices to assist agencies in integrating the Open Data Policy into their operations in furtherance of their missions. The CIO and CTO shall regularly update this online repository as needed to ensure it remains a resource to facilitate the adoption of open data practices.

(b) Within 90 days of the issuance of the Open Data Policy, the Administrator for Federal Procurement Policy, Controller of the Office of Federal Financial Management, CIO, and Administrator of OIRA shall work with the Chief Acquisition Officers Council, Chief Financial Officers Council, Chief Information Officers Council, and Federal Records Council to identify and initiate implementation of measures to support the integration of the Open Data Policy requirements into Federal acquisition and grant-making processes. Such efforts may include developing sample requirements language, grant and contract language, and workforce tools for agency acquisition, grant, and information management and technology professionals.

(c) Within 90 days of the date of this order, the Chief Performance Officer (CPO) shall work with the President's Management Council to establish a Cross-Agency Priority (CAP) Goal to track implementation of the Open Data Policy. The CPO shall work with agencies to set incremental performance goals, ensuring they have metrics and milestones in place to monitor advancement toward the CAP Goal. Progress on these goals shall be analyzed and reviewed by agency leadership, pursuant to the GPRA Modernization Act of 2010 (Public Law 111–352).

(d) Within 180 days of the date of this order, agencies shall report progress on the implementation of the CAP Goal to the CPO. Thereafter, agencies shall report progress quarterly, and as appropriate.

Sec. 4. *General Provisions.* (a) Nothing in this order shall be construed to impair or otherwise affect:

(i) the authority granted by law to an executive department, agency, or the head thereof; or

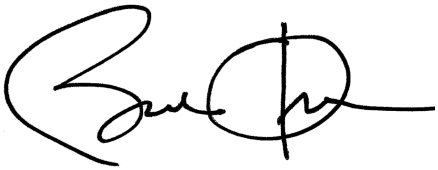
(ii) the functions of the Director of OMB relating to budgetary, administrative, or legislative proposals.

(b) This order shall be implemented consistent with applicable law and subject to the availability of appropriations.

(c) This order is not intended to, and does not, create any right or benefit, substantive or procedural, enforceable at law or in equity by any party against the United States, its departments, agencies, or entities, its officers, employees, or agents, or any other person.

(d) Nothing in this order shall compel or authorize the disclosure of privileged information, law enforcement information, national security infor- mation, personal information, or information the disclosure of which is prohibited by law.

(e) Independent agencies are requested to adhere to this order.



THE WHITE HOUSE,
May 9, 2013.

[FR Doc. 2013–11533
Filed 5-13-13; 8:45 a,]
Billing code 3295-F3

APPENDIX F. JOHN HOPKINS WEBSITE FOR TURBULENCE DATASETS

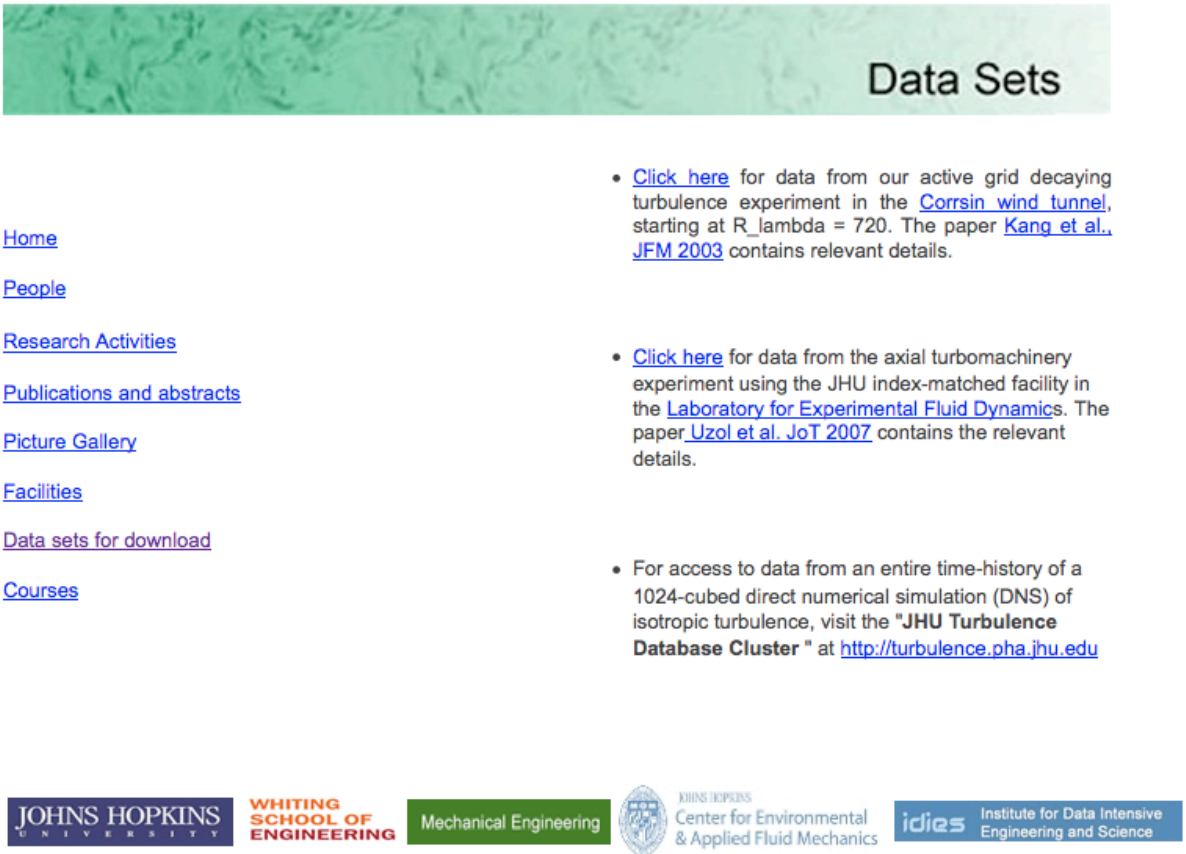


Figure A3.1 Taken from http://www.me.jhu.edu/meneveau/datasets.html

APPENDIX G. GLOSSARY AND ABBREVIATIONS

capability computing	The use of the most powerful supercomputers to solve the largest and most demanding problems, in contrast to capacity supercomputing. The main figure of merit in capability computing is time to solution. In capability computing, a system is often dedicated to running one problem.
capacity computing	The use of smaller and less expensive high-performance systems to run parallel problems with more modest computational requirements, in contrast to capability computing. The main figure of merit in capacity computing is the cost/performance ratio.
confidentiality	A measure of the absence of unauthorized disclosure of data.
conflict	A conflict exists when two or more DOIs have been assigned to the same logical work as represented by identical metadata or when subtle metadata differences may exist (e.g. a slight variation in the publication title or style differences in the item title).
DNS	Direct Numerical Simulation.
DOE	The U.S. Department of Energy.
DOI	Digital Object Identifier.
HPC	High-performance computing. Computing on a high-performance machine. There is no strict definition of high-performance machines, and the threshold for high performance will change over time.
HPCS	High Productivity Computing Systems, a DARPA program started in 2002 to support R&D on a new generation of HPC systems that reduce time to solution by addressing performance, programmability, portability, and robustness.
HSM	Hierarchical Storage Management is a data storage technique which automatically moves data between high-cost and low-cost storage media. HSM systems exist because high-speed storage devices, such as hard disk drive arrays, are more expensive (per byte stored) than slower devices, such as optical discs and magnetic tape drives.
idempotent	An operation that has the same effect whether it is carred out once or multiple times.
integrity	A measure of the absence of improper system alteration.
metadata	Data that describes the organization or contents of other data. The term is used in many contexts, from file systems to large-scale data management.
open source	Software that is available to users in source form and can be used and modified freely. Open source software is often created and maintained through the shared efforts of voluntary communities.
PI	Principle Investigator
Registration agency	Registration Agency:

- Offers services for registration of prefixes and individual DOI names using the DOI system.
- Provides added-value services for registrants and other customers.
- Must be a member of the IDF.
- Engages in marketing, training, development, etc. for their chosen community.
- May maintain a Handle mirror site (optional).
- May subcontract their service provision (optional).

replication The practice of creating and maintaining multiple file copies to ensure availability in the event of hardware failure.

staging Moving data between levels of storage (primary, secondary, and tertiary) for the purpose of managing available space.

tipping point The tipping point is the critical point in an evolving situation that leads to a new and irreversible development. The term is said to have originated in the field of epidemiology when an infectious disease reaches a point beyond any local ability to control it from spreading more widely. A tipping point is often considered to be a turning point.

INTERNAL DISTRIBUTION

1. Terry Jones
2. Sudharshan Vazhkudai
3. Cindy Sonewald, CSR Group Records
4. ORNL Office of Technical Information and Classification

EXTERNAL DISTRIBUTION

4.
5.
6.
7.