
PRACTICE PAPER

Redesigning the DOE Data Explorer to Embed Dataset Relationships at the Point of Search and to Reflect Landing Page Organization

Sara Studwell, Carly Robinson and Jannean Elliott

Office of Scientific and Technical Information, Department of Energy, US

Corresponding author: Sara Studwell (studwells@osti.gov)

Scientific research is producing ever-increasing amounts of data. Organizing and reflecting relationships across data collections, datasets, publications, and other research objects are essential functionalities of the modern science environment, yet challenging to implement. Landing pages are often used for providing ‘big picture’ contextual frameworks for datasets and data collections, and many large-volume data holders are utilizing them in thoughtful, creative ways. The benefits of their organizational efforts, however, are not realized unless the user eventually sees the landing page at the end point of their search. What if that organization and ‘big picture’ context could benefit the user at the beginning of the search? That is a challenging approach, but The Department of Energy’s (DOE) Office of Scientific and Technical Information (OSTI) is redesigning the database functionality of the DOE Data Explorer (DDE) with that goal in mind. Phase I is focused on redesigning the DDE database to leverage relationships between two existing distinct populations in DDE, data Projects and individual Datasets, and then adding a third intermediate population, data Collections. Mapped, structured linkages, designed to show user relationships, will allow users to make informed search choices. These linkages will be sustainable and scalable, created automatically with the use of new metadata fields and existing authorities. Phase II will study selected DOE Data ID Service clients, analyzing how their landing pages are organized, and how that organization might be used to improve DDE search capabilities. At the heart of both phases is the realization that adding more metadata information for cross-referencing may require additional effort for data scientists. OSTI’s approach seeks to leverage existing metadata and landing page intelligence without imposing an additional burden on the data creators.

Keywords: dataset; linked data; related data; landing page; metadata; reuse

Introduction

Scientific research today produces vast quantities of highly complex data, and many data creators face challenges in storing the data and making it accessible, and then in organizing the data in meaningful ways that make it useful for current and future research (MacMillan 2014). These data may already be openly available, but not necessarily organized or linked in relevant and useful ways. Researchers across disciplines echo this sentiment, asserting that ‘. . . the first order problem is not the expression of quality information but rather, finding and linking the disparate pieces of information together to enable the user to make a judgement’ (Blower et al. 2014). The need for more and better linkages across data objects is crucial to provide context to the data and to render it reusable for the consumer. A project using environmental sensing research done at the Center for Embedded Networked Sensing worked to address the challenge of creating better linkages by looking at the lifecycle of the research and capturing artifacts such as instrument data and associated calibration information. These data may have little value in isolation, but high value for researchers trying to reuse the environmental sensing data (Pepe et al. 2009). There are similar difficulties in the biological sciences, as the vast amount of data produced ‘introduces significant obstacles in data organization . . .’ in the endeavor to make the research useful and available to a wider

audience (Chesters, Zhu 2014). Better data organization and integration can be achieved through collaboration with data owners and domain experts, finding ways to exploit existing landing page information and manipulate metadata to create an intuitive search experience. 'For data to be reusable they must be accompanied by a comprehensive description together with . . . discipline-specific metadata to improve data discovery and to facilitate reuse and understanding' (Ulbricht et al. 2016).

Based on experiences and feedback from our data researchers, as well as trends in the data world, we seek to create meaningful contextual relationships among data objects to make the data more accessible by reorganizing DDE. Phase I of the effort to provide additional organization to the DDE database will ensure data relationships are exposed to users. Specific tasks include (discussed in Phase I section):

1. renaming existing collections to more accurately describe their top-level purpose and recasting the data 'Collection' as a second level product type to facilitate grouping of individual Datasets to aid in relational organization;
2. developing mappings to identify vertical and related 'family' relationships between existing records;
3. restructuring of search results display to ensure 'parent' records rise to the top, regardless of user-selected search terms, and that 'child' records continue to behave according to the user-selected sort option;
4. providing linkages from every 'parent' record to the entire 'family' for browsing purposes;
5. adding new records to the appropriate families; and
6. automating the processes involved for sustainability and scalability with the use of new metadata fields and existing authorities.

In Phase II, we will move outside DDE to look at client landing pages to investigate how clients add contextual and organizational features to their landing pages. We will also collaborate with researchers to explore if and how those features can be reflected in the organization of the DDE content. We have learned, through managing the DOE Data ID Service, how challenging it is for busy researchers to provide enough metadata to DDE to make these connections. We will explore how we can apply landing page strategies implemented by data clients into DDE database search and display strategies.

Background

In 2004, OSTI convened a two-day workshop for key staff from DOE data centers to learn more about their challenges and needs. Attendees noted that, while researchers knew where to find necessary data within the boundaries of their own fields, cross-disciplinary research suffered from a lack of knowledge about the data that actually existed across DOE. The need for an inventory of those data holdings was expressed. DDE launched in June 2008 to help researchers and the public discover DOE-funded, publicly available, scientific data. The content was developed by OSTI staff who searched the web to find collections of DOE-funded data, then created descriptive records with hyperlinks to those collections.

Meanwhile, a trend was developing. Could datasets be cited, tracked, and preserved with the same type of scholarly framework as journal publications? OSTI had previously championed the concept of persistent identifiers, obtaining Digital Object Identifiers (DOIs) for journal articles and also assigning and registering DOIs for DOE's technical reports, and was very interested in assigning DOIs to datasets. In 2011, OSTI became a member of DataCite, and began implementation of the DOE Data ID Service to DOE-funded researchers or 'data clients.' Metadata is collected from data clients, and along with a link to each dataset, comprises a dataset record in DDE. As the service began to be better utilized, these 'Dataset' records, with associated DOIs, grew alongside the existing population of OSTI-created 'Data Collection' records. A Data Collection is defined as a record created and curated by OSTI that describes a data project funded by DOE. A Dataset is a single instance of data whose boundaries have been defined by the data client with a DOI associated with the Dataset.

OSTI noticed that some dataset records submitted by data clients stemmed from data collections that also had records in DDE. Unfortunately, the two types of records were not cross-referenced to each other. In addition, the collection records could not include all the terminology found in every 'child' dataset record. The individual dataset records are often very precise and specialized in their metadata, while the metadata of the 'parent' collection record was curated by OSTI and may not share the same terms found in the 'child' dataset record.

The two types of records in DDE formed a co-existence that was ripe with opportunity for the addition of Dataset and Collection linkages. One data client presented a great chance to pilot this idea. The Materials

Project at Lawrence Berkeley National Laboratory began the submission of submitting over 70,000 dataset records. The Materials Project made a request for OSTI to provide a new API that would allow Materials Project researchers to create collections of DOIs to facilitate citation of Materials Project data.

As OSTI began development of this new API, we determined it would be necessary to add a new type of record in DDE – a specialized collection record with an associated DOI. To date, none of the OSTI assigned collections were assigned DOIs. OSTI realized that the necessary work needed to handle this new type of record created an opportunity to better define the parent-child relationships inherent in the two existing types of records in DDE (Data Collections and Datasets). The record types could be better highlighted in DDE, and perhaps other linkages and organization could be added.

Phase I: the Re-envisioning and Implementation of the New DDE Organizational Structure

DDE gains a product type

The original project goal was to capitalize on the natural relationships between Data Collections and individual Datasets by changing the interface for search results to display linkages between the two record types already in the DDE database. Records representing Data Collections would be automatically forced to the top of search results and information to inform users that they are ‘parents’ of Dataset records also returned by the search.

As OSTI began exploring how to create and display relationships between Data Collections and Datasets, it became evident that two record types would not be sufficient to fully realize the hierarchy envisioned by the project team. A third record type was needed to represent a top tier or a level broader than our definition of a Data Collection. The new record type would be the broadest classification, integrating the data client into the definition. As a starting point and for the purposes of this paper, OSTI named this new record type a data ‘Project.’ However, we are having ongoing discussions with our various communities to come up with the most representative name for the data product type. A Project is defined as a collection of data from a specific research group, data center, user facility, or other DOE-funded endeavor. A Project may include multiple Datasets and/or Collections, or may include data from DOE-funded projects without associated DOIs. This definition is similar to the old Data Collection record definition.

Data ‘Collection’ records are now defined as a package of related Datasets, as prescribed by the data client, with a DOI for the entire Collection; the datasets within the collection also have DOIs. Some of the records previously defined as Data Collection records can remain in the revised Collection record category, but many of them would have to be changed to the new, broader Project type. Both Collections and Datasets would have associated DOIs and be curated by the data clients who submit the record.

By adding the Project record type, data clients, creators, owners, and curators can more logically organize their data within DDE, better reflecting what already exists on the dataset landing pages. Context will be added through the creation of top-level Project records that will give an overview of the data produced in that project, and through grouping Datasets into relevant Collections under each Project. Data creators will be able to present their data in a more logical fashion and users will be able to discover and reuse the data through an improved search structure in DDE.

Visually representing the new organizational structure

The new record structure of DDE presents the need to redesign the search interface. To encourage users to search and navigate through records based on the three product types, a more intuitive filtering mechanism that allows users to easily sort from Project to Collection to Datasets is needed.

When a search is completed, the default display will be the Project records. Collection and Dataset records will be viewable by navigating to the respective filter option tabs available at the top of the results (see **Figure 1**). This mimics most landing page designs and navigation, in that the user is presented with more general information before drilling down to specific data, and provides context for the data.

Search and navigation using the new relations

Additional information and navigation options will be added to the search results (see **Figure 1**) and to the individual citation pages (see **Figure 2**) to enhance the ability to browse among linked records. Listed under each Project result record’s metadata is the number of associated Collections and Datasets (or for Collections, the number of associated Projects and Datasets). Selecting any of these takes the user to a results list for the respective record type. Similar navigation options were added to individual citation pages. The user has the option to toggle among the Project and associated Collections and Datasets either by using the tabs at the top of the page or by following the hyperlinks in the right-hand menu.

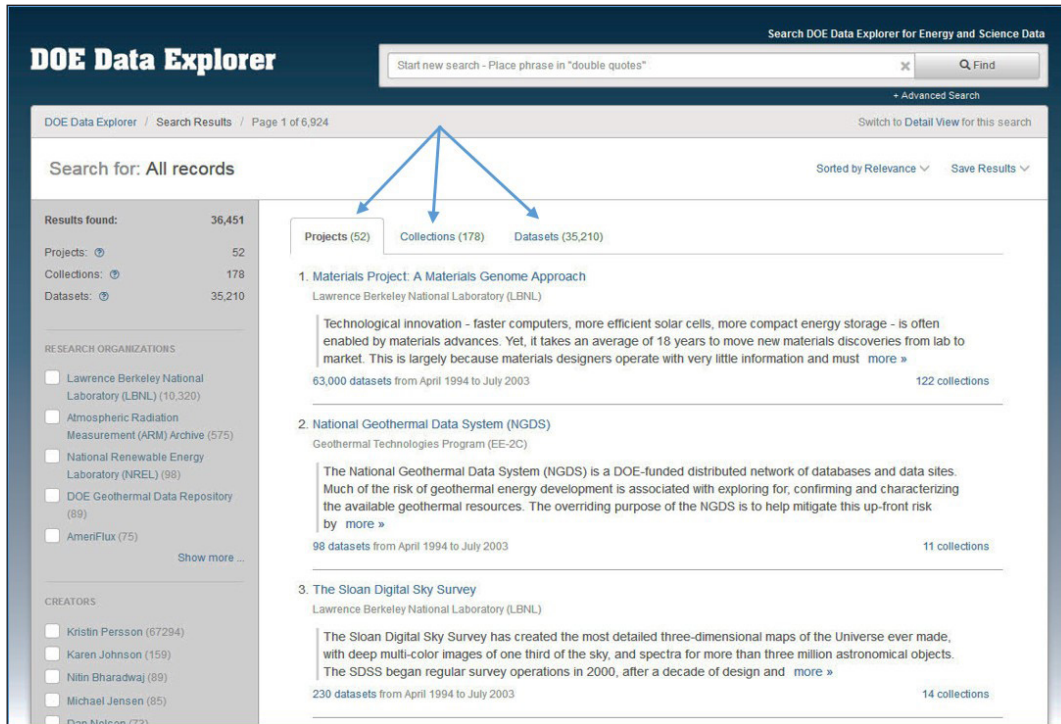


Figure 1: Default results view with product type filtering and navigation.



Figure 2: Individual citation page with associated product type navigation.

This reinvention of DDE presented the opportunity for more enhancements to improve search experience and data discoverability. Offering additional filter options on the citation page allows users to discover additional Projects (or Collections or Datasets) that they may not otherwise find. ‘More Like This’ features are commonplace in many commercial search engines and can also be very useful in STI collections. One of OSTI’s clients, the Geothermal Data Repository (GDR), provides links to related datasets on each record page. We are adding the ‘More Like This’ capability to DDE; with each search, a list of suggested related records will appear in the right hand menu created by the related terms in each of the metadata records. We run a search across a specific set of metadata fields, such as the research organization and subject fields, identify matches or related terms, and display those records in the ‘More Like This’ section. A list of journal articles or other scholarly works that cite the Collection or Dataset DOI can be accessed either by selecting the ‘Cited by’ tab at the top of the page or by selecting the hyperlinked number listed in the right-hand menu. Building on citation navigation in some of OSTI’s other discovery tools, DDE will be implementing this using citation information from Thompson Reuter’s Data Citation Index.

We also recognized the value of being able to view information about related Datasets within the context of the overarching Project, so the option to see additional metadata for a related Dataset will be available via an expandable ‘Details’ view for each record on the Datasets tab (as shown in **Figure 3**). This allows the user to determine, without leaving the overarching Project or Collection record, if the related Dataset is relevant to his or her search. This ‘Details’ feature will also be implemented on the general search results page. The ability to determine whether the data is relevant to the user quickly through visual cues is also valuable. We are adding a selection of illustrative images or charts from the Data Project record (or Collection or individual Dataset) to each citation page, giving the user another feature to help them discern the value of the record.

DOE Data ID Service workflow

Metadata for individual datasets is submitted via the DOE Data ID Service to OSTI’s input repository and processing system, known as Energy Link (E-Link), by the data owner, either on an individual basis or via an API. Once the metadata has been validated and assigned a DOI, it is sent to DataCite, and then published as a record in DDE (see **Figure 4**). No linkages between the existing Collections and Datasets exist, but

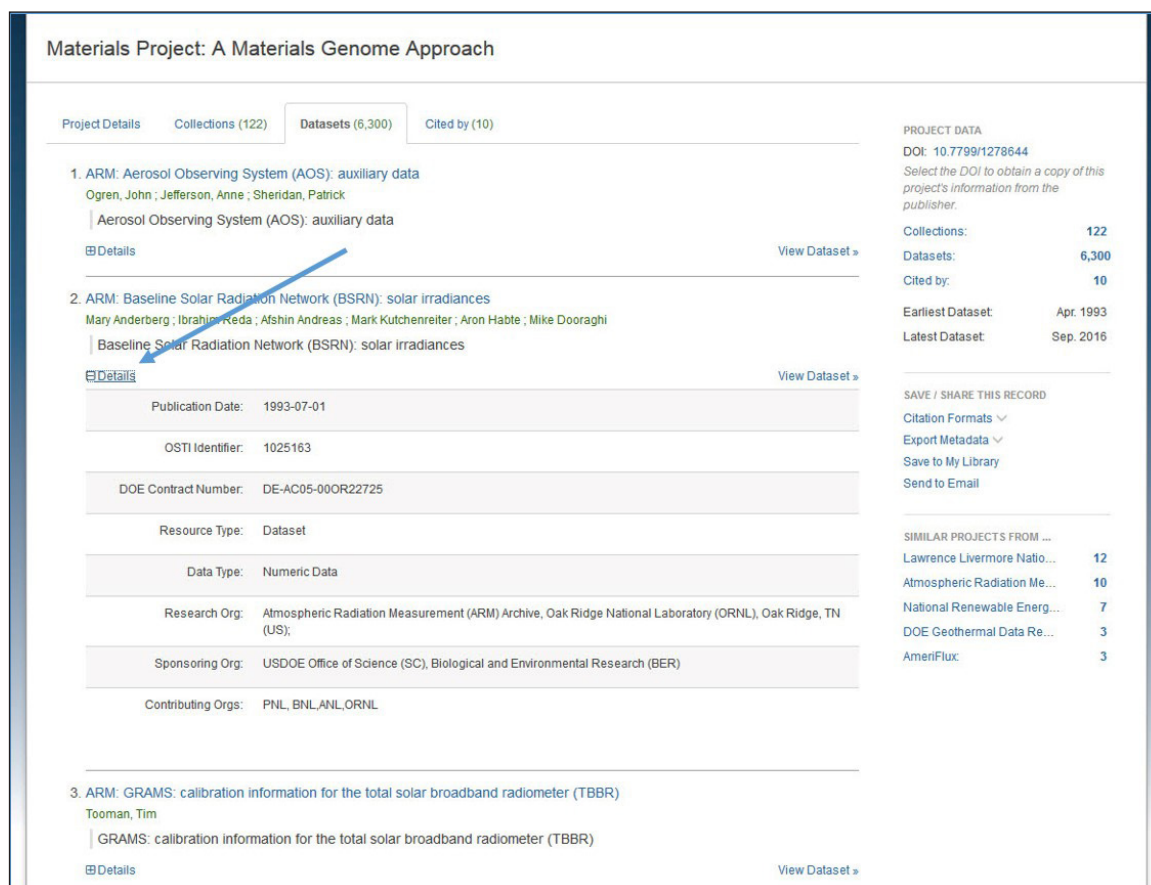


Figure 3: Additional metadata details for related Datasets.

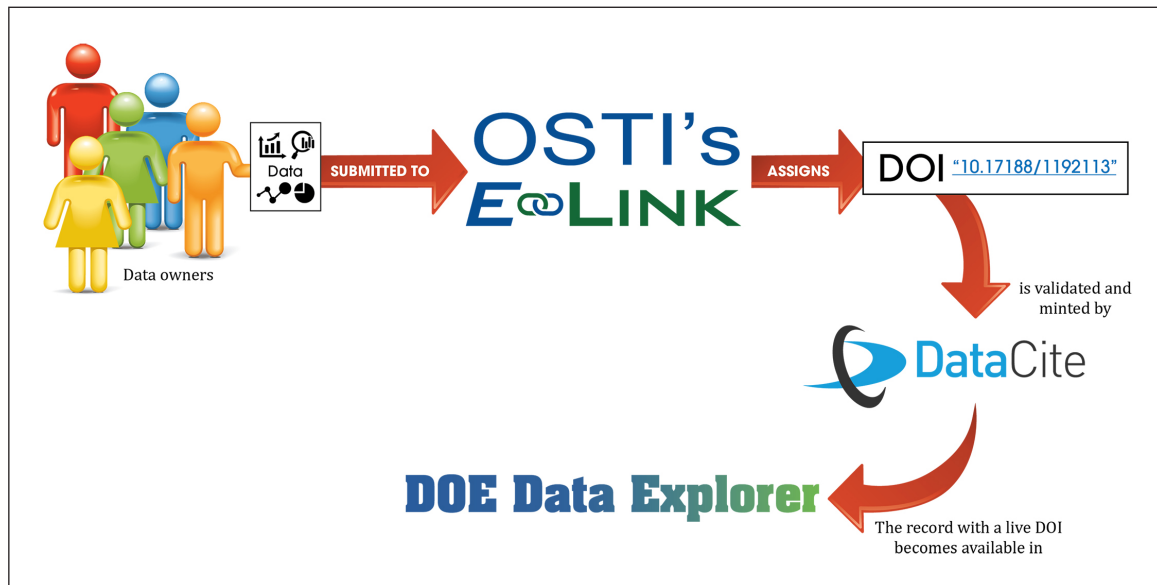


Figure 4: Data ingest, DOI assignment, and publication.

support for this has been accomplished using OSTI's existing technical infrastructure for STI. The addition of specific referential metadata facilitates easy establishment of relationships among records, and the linkages can then be added. Collection and Dataset records will be created, assigned DOIs through DataCite, and then related to one another appropriately using the new functionality. These hierarchical relations will be manifested in DDE.

Phase II: Collaborating with Clients to Create More Robust Associations *Using landing pages as guides for organizing data in DDE*

Phase II of the effort will investigate how clients add contextual and organizational features to their landing pages and websites. With researcher collaboration, we will explore how those features might be reflected in the organization of the DDE content, particularly with utilization of the new Collections product type. The Materials Project at Lawrence Berkeley National Lab is attempting to form parent/child relationships through the bundling of DOIs, creating Collections. In this case, the submitting data client is doing the work of exposing important relationships to users at the point of search. Following this model, OSTI will allow the data owners to define Collections based on the configuration already found on the website where the data is hosted. This linkage provided by the data owners also provides DDE with a relational framework between individual Datasets and an overarching Collection that groups datasets in a meaningful way.

The Atmospheric Radiation Measurement (ARM) Archive tackled DOI assignment granularity and the linkage between datasets for their instrument datastreams in a different way, assigning one DOI for each instrument that continuously collects data. The metadata in DDE for each ARM Dataset is for a specific instrument and the link to the dataset is to the instrument landing page. The ARM data repository provides each user with a specific citation, including timestamps and additional metadata, generated when the user downloads a dataset from the repository. ARM has created relationships between the DOI of the instrument's datastream and the unique citations generated each time slices of the data are obtained. Like the work done by Pepe et al. (2009), working with ARM data owners, we could consider leveraging this organizational structure in DDE, creating Collection records for each instrument. Datastreams from each instrument could become Dataset records and have individual DOIs. This is worth investigating as it may help the DDE user better understand the data produced by ARM, how much data is available, and the context in which it is collected (in this instance by location and instrument).

Another of OSTI's data clients, The DOE Geothermal Data Repository (GDR), provides links to related datasets on its dataset landing pages. This allows users to discover important background information about how the data is collected and provides insight into a larger scope of data that may be relevant to the user. We already have metadata records for the datasets accessible through the GDR, but there is no information available about how they are related to one another. We will explore leveraging these existing

associations provided on the dataset landing pages on a broader scale, working with the data producers and owners to group related datasets into Collections so that DDE users can discover more information at the point of search.

Lessons learned and additional areas for exploration

Phase I of this work introduced a number of lessons learned that will help moving into Phase II. For example, manual curation of the existing data Collections into Projects and the subsequent initial plans for mapping related records have been the biggest challenge. This challenge mirrors what previous research has found regarding the difficulty of organizing data in ways that make is useful for others (MacMillan 2014). We have determined that collaborating with our data clients and depending on their domain-specific knowledge in creating data records and defining subsequent linkages is the best solution, echoed by Pepe et al. (2009) in previous research.

From these lessons learned we have uncovered a number of questions and considerations. Some of these include:

1. Other clients are focusing on tying a digital knot between datasets, publications, and related research objects such as software. How can we start to better interlink data, software, and publications to provide a more comprehensive research environment?
2. How do we create additional hierarchical associations such as “lab rollups,” associating user facilities (like Argonne National Lab-Advanced Photon Source) to the overarching lab (Argonne National Lab) so that a user can find data related to an instrument instead of an individual project? This is already evident in ARM records, where there exists one overarching Project record for ARM, and each of the instruments will have a Collection record, with temporal considerations addressed at the Dataset level.
3. Some data can logically be related to more than one Project. How can we address these linkages and expose them in DDE?

We are currently in Phase I of the project, having created the internal architecture necessary to ingest and relate the types of data. Phase I completion is estimated by the end of 2017, when Phase II will begin to work through the questions listed above. As we move forward with the multi-phase restructuring of DDE and corresponding ingest infrastructure modifications, we will work closely with our data clients and interested stakeholders (such as data owners and producers) to ensure that their concerns are addressed and that inherent, logical relationships are clearly reflected in the data. OSTI’s goal is to make data more accessible and understandable, and to help users see contextual relationships early in the search experience.

Competing Interests

The authors have no competing interests to declare.


References

- Blower, J, Lawrence, B, Kershaw, P and Nagni, M** 2014 CHARMe Commentary Metadata for Climate Science: Collecting, Linking and Sharing User Feedback on Climate Datasets. *Geophysical Research Abstracts*. In: *Proc. of EGU General Assembly 2014*. Vienna, Austria on 27 April–02 May 2014.
- Chesters, D and Zhu, C-D** 2014 A Protocol for Species Delineation of Public DNA Databases, Applied to the Insecta. *Systematic Biology*, 63(5): 712–725. DOI: <https://doi.org/10.1093/sysbio/syu038>
- Macmillan, D** 2014 Data Sharing and Discovery: What Librarians Need to Know. *The Journal of Academic Librarianship*, 40(5): 541–549. DOI: <https://doi.org/10.1016/j.acalib.2014.06.011>
- Pepe, A, Mayernik, M, Borgman, C L and Van de Sompel, H** 2009 From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web. *Journal of the Association for Information Science and Technology*, 61(3): 567–582. DOI: <https://doi.org/10.1002/asi.21263>
- Ulbricht, D, Elger, K, Bertelmann, R and Klump, J** 2016 PanMetaDocs, ESciDoc, and DOIDB—An Infrastructure for the Curation and Publication of File-Based Datasets for GFZ Data Services. *ISPRS International Journal of Geo-Information IJGI*, 5(3): 25. DOI: <https://doi.org/10.3390/ijgi5030025>

How to cite this article: Studwell, S, Robinson, C and Elliott, J 2017 Redesigning the DOE Data Explorer to Embed Dataset Relationships at the Point of Search and to Reflect Landing Page Organization. *Data Science Journal*, 16: 17, pp. 1–8, DOI: <https://doi.org/10.5334/dsj-2017-017>

Submitted: 25 October 2016 **Accepted:** 15 March 2017 **Published:** 04 April 2017

Copyright: © 2017 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS 