Article

panMetaDocs, eSciDoc, and DOIDB—An Infrastructure for the Curation and Publication of File-Based Datasets for GFZ Data Services

Damian Ulbricht [1],*, Kirsten Elger [1],*, Roland Bertelmann [1] and Jens Klump [1,2]

[1] GFZ German Research Centre for Geosciences, Telegrafenberg, Potsdam 14473, Germany;
    roland.bertelmann@gfz-potsdam.de (R.B.); jens.klump@csiro.au (J.K.)

[2] Commonwealth Scientific and Industrial Research Organisation, Mineral Resources, 26 Dick Perry Avenue,
    Kensington, WA 6151, Australia

* Correspondence: damian.ulbricht@gfz-potsdam.de (D.U.); kirsten.elger@gfz-potsdam.de (K.E.);
    Tel.: +49-331-288-2392 (D.U.); +49-331-288-2822 (K.E.)

Abstract: The GFZ German Research Centre for Geosciences is the national laboratory for Geosciences in Germany. As part of the Helmholtz Association, providing and maintaining large-scale scientific infrastructures are an essential part of GFZ activities. This includes the generation of significant volumes and numbers of research data, which subsequently become source materials for data publications. The development and maintenance of data systems is a key component of GFZ Data Services to support state-of-the-art research. A challenge lies not only in the diversity of scientific subjects and communities, but also in different types and manifestations of how data are managed by research groups and individual scientists. The data repository of GFZ Data Services provides a flexible IT infrastructure for data storage and publication, including minting of digital object identifiers (DOI). It was built as a modular system of several independent software components linked together through Application Programming Interfaces (APIs) provided by the eSciDoc framework. Principal application software are panMetaDocs for data management and DOIDB for logging and moderating data publications activities. Wherever possible, existing software solutions were integrated or adapted. A summary of our experiences made in operating this service is given. Data are described through comprehensive landing pages and supplementary documents, like journal articles or data reports, thus augmenting the scientific usability of the service.

Keywords: data publication; data curation; persistent identifier; institutional repository

1. Introduction

The free and open access to research data has been identified as a key issue by the scientific community, by research agencies and governments (e.g., Berlin Declaration, G8 Science Ministers Statement, EU Implementation of the G8 Open Data Charter, Digital Agenda of the Federal Government of Germany [1–4]). As a consequence, there is a dynamic co-evolution of national and international guidelines on the management of and on open access to research data (e.g., Berlin Declaration, OECD Declaration, INSPIRE Directives [1,5,6]). In parallel the scientific community developed technical and conceptual models to make research data citable and usable by others. In this context, Digital Object Identifiers (DOI) have emerged as the leading identifier system for text and data publications [7].

Storage in appropriate, discipline specific repositories is an essential part of ensuring access and long-term preservation of these data. For data to be reusable they must be accompanied by a comprehensive description together with standardized and discipline-specific metadata to improve data discovery and to facilitate reuse and understanding. Publishers, data publishers, and domain

scientists have committed themselves to guidelines on how to embed links to datasets in publications, and corresponding links from data facilities to journals, through persistent identifiers. Datasets should ideally be referenced by persistent and globally unique, resolvable identifiers, such as Digital Object Identifiers (DOI) [7]. This article is primarily directed at data curators, repository managers, and data scientists, but may also be of interest for researchers, especially the sections about landing pages and data description (Sections 3–5). After a general overview, we explain the design and functionality of each software component (eSciDoc, panMetaDocs, DOIDB), give a brief overview of different formats for data publication, the introduction of the metadata editor, examples of DOI landing pages, licences of the software components, and conclude with a discussion and outlook section.

The Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences is the national laboratory for geosciences in Germany and member of the Helmholtz Association, Germany's largest research organisation. The focus of GFZ is to study the history of the Earth and its characteristics, the processes occurring on its surface and within its interior, as well as the interactions between the geosphere, the hydrosphere, the atmosphere, and the biosphere. The resulting research data cover all geoscientific disciplines. They range from large dynamic datasets derived from global monitoring seismic, magnetic, or geodetic networks with real-time data acquisition or climate station data, to remotely-sensed satellite products, to various model results, to geochemical analyses from various labs, and field observations. In the geosciences, the overall management and curation of data is challenging. Sizes of datasets vary from rather tiny to quite large. There are a variety of processing software solutions for data and a resulting bunch of file formats and metadata schemas. Often the acquisition of data is expensive and often the research data record a natural phenomenon that can be observed only once. All these data are valuable data and deserve to be preserved for future generations of scientists [8].

As part of the scientific research, significant volumes and numbers of research data are generated and subsequently become source materials for publications. The development and maintenance of data systems is an essential pillar of GFZ activities to support state-of-the-art research. Large dynamic datasets from global monitoring networks produce real-time or near real-time data streams that are fed into data repositories and are made accessible through international theme-specific data portals. Examples at GFZ for this are systems like GEOFON, GNSS services, INTERMAGNET, and many more. GEOFON [9] operates a global seismic monitoring network and an international data centre and archive which facilitates real-time access to seismological data. As part of the International GNSS Service IGS [10], GFZ operates around 30 globally distributed multi-GNSS reference stations and hosts an analysis centre for GNSS data. The 18 magnetic observatories of GFZ and cooperation partners are part of the International real-time magnetic observatory network INTERMAGNET [11]. Besides such data from monitoring networks there is a "long tail" of highly specified and very variable research data of individual scientists or research groups which also need to be curated and published.

GFZ Data Services is the central service point for research data, in particular for publication and archiving of data. GFZ Data Services are designed to serve the data management needs of researchers, following extensive consultations with relevant stakeholders. In addition, GFZ Data Services developed and maintain a data repository and offer publication of datasets and data products in conjunction with DOI registration. DOI registration is also offered as a service to other repositories at GFZ, e.g., GEOFON. The datasets published at GFZ cover the full range of data publications as described in Katz and Strasser (2015) [12]: data may be published as supplementary material to journal articles, with a descriptive article in one of the new "Data Journals", or as independent entities. To provide a format for extensive description of independently published datasets, GFZ Data Services launched a data report series in 2011. This wide range of data publication clients requires the system to be able to handle various different metadata schemas and to offer flexible and fine-grained management of access rights to the system and its holdings.

Especially with respect to the increasing expectations and requests for open access to research data, the conservation and curation of research datasets and the maintenance of bespoke data curation

software beyond the lifetime and funding of a project are key tasks for GFZ Data Services. To meet these demands, the repository of GFZ Data Services was built as a modular system to allow an easy adaptation to changing requirements, to be able to handle the high variability in data types, and also be suitable for the large variety of geoscientific disciplines present at GFZ. Wherever possible, existing and proven software solutions were integrated or adapted. The package roughly consists of the framework eSciDoc, the essential repository at the base, panMetaDocs as application to publish datasets, and DOIDB as a facility to register DOI and metadata. This modular approach also allows us to change parts of the service over time without the need of a complete re-engineering of the system (see Figure 1 for the general structure of the GFZ Data Services Repository). The data publishing services are closely connected to PubMan, an application for textual publication [13] that is also based on eSciDoc and therefore part of the well-established publishing services of the library of GFZ. The library acts also as the central library for four research institutes in Potsdam (Library Wissenschaftspark Albert Einstein Potsdam).
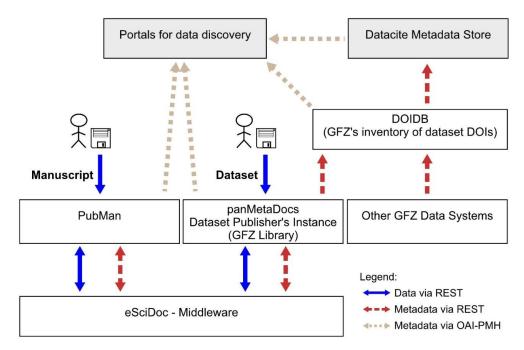

Figure 1. Illustration of the system architecture of the modular software assemblage for the repository of GFZ Data Services. Grey boxes indicate systems in the public domain.

The curation policy of GFZ Data Services is based on a separation of concerns between research project and memory institution (e.g., library). In this concept, the data curation continuum from data generation through data storage to data access is divided into four "Domains of Responsibility" [14,15]. These "Domains of Responsibility" in research data management help to delineate the responsibilities of the actors involved. They also outline the context of shared knowledge about data and, in this way, help to determine how detailed descriptive metadata need to be in order to allow the reuse of research data and at which point implicit contextual metadata need to be encoded into the stored metadata. GFZ Data Services operate in the "Persistent" and "Access" domains and work with researchers to transfer their data into the "Persistent" domain for publication and archiving. More detail on how we implemented the curation domains based on eSciDoc can be found in Klump *et al.* (2015) [14].

2. Software Components for Data Publication

Our system for data publication makes use of four software components that are combined in a modular fashion. We collect research articles, datasets and associated metadata with the web-based software components PubMan for articles and panMetaDocs for datasets. Storage of these digital

objects is mediated by an eSciDoc repository middleware. DOIDB, our fourth component, serves for minting dataset DOIs through the DataCite DOI registration service at the German National Library for Science and Technology (TIB Hannover) and acts as metadata portal for datasets published through GFZ Data Services. The software components are coupled through web services. In addition, metadata can be harvested from each component into overarching metadata portals, e.g., B2FIND or Earth Cube, by using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [16].

Figure 1 outlines the architecture of the publishing system at GFZ. Arrows indicate the flow of data or metadata and the protocols used. eSciDoc, as a middleware, does not provide graphical user interfaces. It has a representational state transfer interface (REST) [17] and its entire functionality is provided through web service endpoints. The applications PubMan and panMetaDocs provide graphical user interfaces for annotation of data files with metadata and to store data and metadata in eSciDoc using the eSciDoc web services. Data publication by GFZ, as a DataCite publication agent, is also available to data systems outside of the eSciDoc framework. These publications may be registered through DOIDB. To monitor and moderate DOI registration from GFZ data systems, we modified the source code of the DataCite Metadata Store to act as a proxy to the DataCite DOI registration infrastructure and published the software as DOIDB. The functionality of DOIDB will be discussed in more detail later in this paper. The metadata can be used by portals to navigate datasets and publications, to show a rough overview of the content, and to offer links to landing pages. The landing pages describe data and publications in further detail, contain links to related material, and provide files for download.

All software components are based on free and open source software solutions and have been modified to fit our specific needs. In particular, we modified panMetaWorks [18] to make use of eSciDoc for data storage, instead of a file system, and published the newly derived software as panMetaDocs [19,20]. This software solution will also be described in more detail in the course of this paper.

The data publication workflow offers the possibility to have an intermediate state during which the datasets and metadata are not yet published to the public domain, but at the same time fully accessible to, e.g., the reviewers of a journal article about the datasets. This intermediate state is also used to coordinate details about the datasets and metadata with the data creator.

2.1. eSciDoc—The Implementation of eSciDoc as Repository Infrastructure at GFZ

eSciDoc was jointly developed by the Max Planck Society and FIZ Karlsruhe—Leibniz Institute for Information Infrastructure. Funded by the German Federal Ministry of Education and Research (BMBF), the development of eSciDoc was inspired by the requirement to build a framework for virtual research environments for multi-disciplinary research organisations. eSciDoc is a middleware for internet based research applications [21]. It implements functions commonly used for assembling and publishing binary files, such as creating, reading, updating and deleting. Besides these fundamental services, eSciDoc offers a range of additional services, one of them being authentication and authorisation of users.

At GFZ we established eSciDoc as institutional storage middleware for both, textual and research data publications. As mentioned in the introduction, this was done primarily with the aim to simplify the data management by providing a common storage infrastructure at GFZ that is independent from research project specific data management solutions. Project-specific software can only be maintained while a research project is active. After the end of a research project, it becomes unsustainable to maintain dozens of project-specific software instances to ensure continued access to the stored data. eSciDoc addresses this maintenance and access challenge by decoupling data curation applications from data storage by providing an Application Programming Interface (API) to a common data storage infrastructure.

In our applications we make use of eSciDoc "items", which are the basic entities of the eSciDoc information model. The eSciDoc content model also allows for data to be stored in more than one

representation form described by more than one metadata schema [14,19]. eSciDoc items may also consist of multiple files, which again can be described by an arbitrary number of metadata records. Updating an eSciDoc item automatically creates a new version of the item.

All access rights for data objects administered through eSciDoc, including the publication process, are defined in eSciDoc. By setting appropriate access rights, the system allows to declare datasets as private, shared within groups, or published and accessible to the general public. This applies to all applications accessing data objects through eSciDoc, e.g., panMetaDocs and PubMan. Published files can be publicly accessible, or can require user authentication and authorisation. The latter may apply in the case of embargo periods on data release, or to guard against accidental download of very large datasets. Supported authentication protocols are Lightweight Directory Access Protocol (LDAP), Shibboleth and OpenID [22–24]. The system at GFZ Data Services uses LDAP to be able to reuse user information maintained by the centralised identity management facilities supported at GFZ. As this limits access to the system to GFZ employees, we also describe our solutions to expose datasets to external reviewers and ways to collect metadata from external scientists later in this article.

2.2. panMetaDocs—Describe and Publish Research Data

panMetaWorks, from which panMetaDocs [25] was derived, is a web-based, collaborative, metadata and data exchange platform for distributed research projects developed by Robert Huber at PANGAEA [18]. The aim was to provide an easy-to-use graphical user interface, to offer a rich metadata editor, and to facilitate the dissemination of metadata through Really Simple Syndication (RSS) and OAI-PMH interfaces. Access to data and metadata can be set private to the individual researcher, shared within a project group, or publicly available on the internet. In a situation where data curation is distributed between several groups, the OAI-PMH interface of panMetaWorks enables the harvesting of metadata to a central data catalogue of the institution or project.

Since panMetaWorks provided a rich set of features that would be useful to research projects at GFZ, we decided to adapt panMetaWorks to our requirements. We obtained permission to reuse the source code of panMetaWorks and modified it to use the eSciDoc API for data storage and user authentication and to make the administration of diverse metadata easier. panMetaDocs installations are able to store metadata in various formats and commonly used schemas as Dublin Core, DataCite [26], the INSPIRE profile of ISO19139 [27], and NASA GCMD DIF [28]. Storing metadata in multiple formats with an eSciDoc item allows the description of datasets in more facets than those provided by only one common denominator. This is required to handle the broad range of data produced at GFZ. In addition, metadata records can be enriched over time as the data object is transferred from one domain of responsibility to another [14]. Starting from the panMetaWorks code base, we revised the metadata editor and changed the user interface from HTML4 to HTML5. The metadata editor is now a Javascript application using the Sencha ExtJS framework to render elements of the graphical user interface, to parse Extensible Markup Language (XML), and to generate XML. The functionality of the metadata editor will be described in more detail later in this paper. To allow data sharing among project partners, panMetaDocs instances have their own interfaces for metadata dissemination to bypass access rules of eSciDoc that do not allow the dissemination of metadata of unpublished datasets.

To grant reviewers access to a non-published dataset independent from the authentication system of eSciDoc, we create a cryptic temporary link leading to a preview of the future landing page. The preview includes all data and metadata with the correct future DOI that may be evaluated during the peer-review and the correct citation can already be added to the article's reference list. In addition, the review link turned out to be useful to show scientists a preview of their datasets during the publication process. To prevent scientists from using the DOI before its registration, the preview page contains a note that the dataset is currently under review.

2.3. The DOIDB—A Proxy DOI Minting Agent

GFZ became a "publication agent" for data publication in the context of the project "Publication and Citation of Primary Scientific Data" (STD-DOI) [29,30]. In this precursor to DataCite, the international organisation that operates the DOI service for data publication today, the German National Library for Science and Technology (TIB Hannover), delegated data publication and storage to "publication agents" which registered their data publications with the DOI registration service at the TIB Hannover. While the other project partners in STD-DOI operated single data repositories, GFZ had to integrate a number of pre-existing repositories (e.g., GEOFON) to use this service. To log and moderate DOI registrations at GFZ we developed a DOI registration proxy service, which we called DOIDB.

With the foundation of DataCite TIB Hannover became a DataCite "allocation agent". Its DOI registration service for data publications received a number of technical updates, one of them being the metadata store and DOI minting service. In October 2010, DataCite released the source code of the new component, called DataCite Metadata Store (MDS), on GitHub. In addition to the DOI registration, the MDS has a component for searching datasets by metadata attributes and it provides a component for the dissemination of its metadata catalogue through an OAI-PMH interface. The MDS also manages the DOI registration accounts for the associated data centres using this service to mint dataset DOIs.

The changes to the DataCite API at central level also called for changes in our DOIDB. Instead of changing the original software at GFZ, we decided to reuse the DataCite MDS and adapt it with minor modifications [31]. Like in its previous function, the new DOIDB acts as a proxy service between GFZ data systems and the DataCite Metadata Store. Besides the DataCite Metadata schema [32], DOIDB accepts two additional schemas specific to the geosciences (ISO19139 [27] and NASA GCMD DIF [28]). We published the DOIDB source code on GitHub [33–35] and we use git mechanisms to be synchronous with current developments on the original code base from DataCite.

3. Formats for Data Publication

High-quality research data must be accompanied by metadata ensuring that data formats are documented and understandable to allow reuse by others. Assuming the scientifically correct and careful acquisition and processing of a dataset, the scientists, the data repository or the publisher should make sure that the dataset is accompanied not only by standardised metadata for data discovery, but also by an adequate and sufficient description suitable for data reuse (*i.e.*, structural metadata [36]).

Data supplements to scientific articles form the largest number of data publications at GFZ so far. It is increasingly accepted and even promoted by publishers to store supplementary data to an article in a discipline-specific or institutional data repository, and not directly attached to the article [7]. An advantage of using an open access data repository like the one described here, is that datasets are freely available even when the journal is not an open access journal. Furthermore, supplements will not get lost if the journal moves to another publisher, as it has happened in the past. We recommend scientists to send their manuscripts to a journal and publish the supplementary datasets in the GFZ Data Repository. Technically, we use panMetaDocs to ingest data and metadata into eSciDoc. In the case of data accompanying a publication, the dataset may be published at the same time as the scientific articles (if the data is not published before anyway), both parts are cross-referenced, and the dataset needs to be accessible for the peer-review process. In this case, we provide temporary access links for the review process, as described in Section 2.2, and register the DOI at a later stage, to be synchronous with the publication of the article.

Data reports, published at GFZ, have proven to be suitable publication formats for comprehensive descriptions of published datasets, in particular for independently-published datasets. These data publications are reviewed in an internal review process and published in a standardised format whenever possible. Data reports are published together with the datasets. In this case, we use PubMan and panMetaDocs independently to put a manuscript and research data into eSciDoc and publish both

entities. The manuscript and the datasets both receive individual DOIs that are cross-referenced via DataCite metadata.

For large datasets that are stored in international data centres, or that are too large to be accessed through a landing page, we provide panMetaDocs instances that only collect metadata to be used for the DOI landing page. In this case, the download links to the data files may be directed to an external server.

Due to the diversity of research disciplines at GFZ, the file-based repository framework is open to store data in any format as soon as they are provided as individual files. Nevertheless, scientists are encouraged to deposit files in formats that are recommended for preservation by the Library of Congress [37]. In exceptional cases file formats popular in the respective community are accepted.

4. User Interface for Metadata Assessment

To fill the metadata schemas with valuable information both static information generated from the project context and dynamic information entered by scientists are used. The separation in static and dynamic content is required to generate easy-to-use forms that do not bother users by forcing to enter the same information repeatedly.
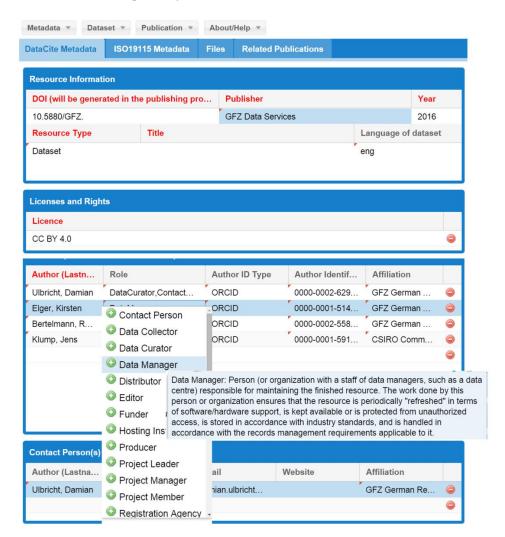

Figure 2. Extract of the graphical user interface (GUI) of the Javascript-based metadata editor for the collection of metadata and digital object identifiers (DOI) registration, based on the DataCite 3.1 metadata schema [32] and complemented by theme-specific fields as required for GFZ data. Mandatory and recommended metadata fields are colour-coded (red letters = mandatory fields). Whenever possible, nominal lists and explanations in pop-up windows were implemented.

Figure 2 shows the graphical user interface (GUI) of the JavaScript metadata editor [38] that has been developed to create standardised metadata for data discovery, which is the base for the DOI registration. The scientist may enter the metadata into the form, save an XML file on its local drive, load XML files into the editor, and submit the final version to GFZ Data Services. Here, additional functions, like the DOI registration and synchronisation with panMetaDocs and eSciDoc, are possible after a review of the metadata. The metadata editor is a stand-alone application that can be used without panMetaDocs or eSciDoc. This allows the submission of metadata by the scientists themselves. In addition, external users, e.g., project partners, needed a way to provide metadata without logon to the eSciDoc system.

To increase the usability of the metadata editor, the metadata fields shown in the GUI aim to be usable by and understandable for scientists. This includes not requesting manual entry for information that may be retrieved automatically, e.g., the URL of a specific Creative Commons Licence, but also the renaming of certain fields in the "scientific" language, e.g., "creators" of DataCite are called "authors". This adaptation was made only in the GUI; the database itself keeps the required terms as required by supported metadata standards. To avoid typographical errors, drop down menus are used and explanations to possible entries are provided via pop-up windows (see Figure 2). A full documentation of metadata fields and the usage of the metadata editor is available for download in the help section. As an additional service, the metadata editor is equipped with an interactive mapping tool through which geographic coordinates may be retrieved from the map and automatically entered in the respective metadata fields, but also serves as visual feedback for manually-entered coordinates that are immediately displayed in the map. Multiple geospatial entries as boxes or points are possible.

5. Landing Pages—Discipline-Specific Presentation of Datasets

Figure 3 shows a DOI landing page generated by XSLT stylesheet transformation of an eSciDoc item containing the metadata generated with panMetaDocs. Since every metadata schema has its strengths and weaknesses, the visual presentation in an internet browser is a mixture of the information provided from eSciDoc and from the information scattered over the different metadata schemas. The information about download restrictions for an embargo was stored in eSciDoc, the contact information and spatial coverage were stored in the ISO19139 schema [27], and the science keywords come from the NASA GCMD DIF schema [28]. Furthermore, the DataCite Metadata Schema [32] was used to link to different kind of publications. The citation information of the linked publications (entered with the DOI only in the metadata form) is downloaded from DataCite and Crossref at time of display in the browser using Javascript.

The clear and user-friendly presentation of datasets is an additional challenge for the creation of landing pages, as generated data differs not only in file formats but also most notably in the research-domain specific metadata for reuse. This is especially relevant for the highly diverse GFZ data of different scientific disciplines, and led to the development of discipline-specific DOI landing pages. This is mainly achieved by the reduction of information to the relevant elements for each discipline and is improving the usability and re-use of published datasets. A global gravitational model published via the International Centre for Global Earth Models [39], e.g., does not require a map (Figure 4), whereas maps with multiple bounding boxes are essential for the presentation of hyperspectral flight campaigns with associated field observations within the German Environmental Mapping and Analysis Program (EnMAP [40], Figure 4). Nevertheless, the standardised metadata for data discovery is always collected and available for download and inline view on the landing page (Figure 3C).

An additional advantage of having different types of landing pages is the possibility to create project-specific designs for large collaborative projects between GFZ and external partners (Figure 4). By generating scripts to present data in a community specific way, it is possible to select the HTML rendering at time of minting the DOI.
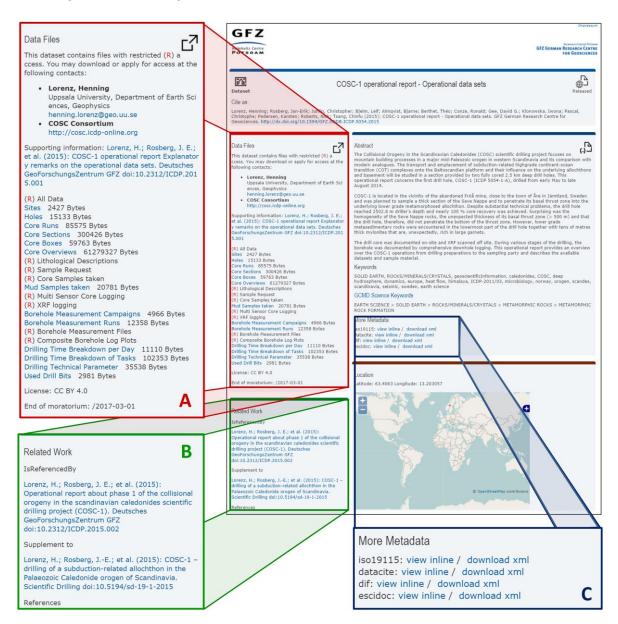

Figure 3. A DOI landing page generated by XSLT stylesheet transformation of an eSciDoc item (example from the International Continental Drilling Program [41]). Highlighted are information about the data files, contacts, and download restrictions (**A**), linked articles and linked datasets (**B**), and download links for the metadata in different metadata schemas (**C**). In this example, the data files are partly open access and directly downloadable via the landing page (file size is indicated), while other data files have restricted access for a defined embargo period (marked red R). The map data are © OpenStreetMap contributors and published under a CC-BY-SA license.
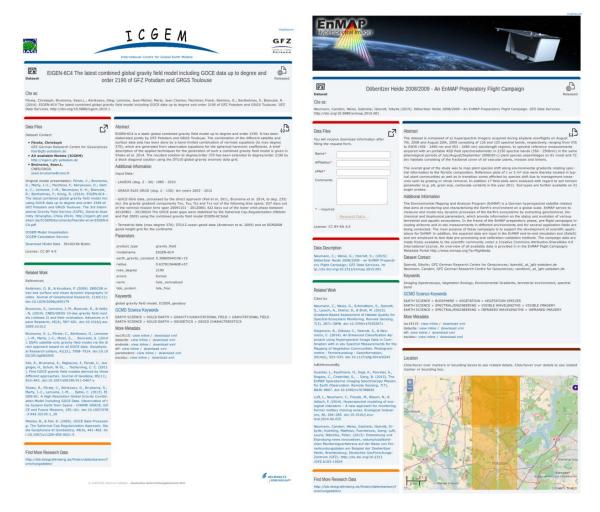
Figure 4. Different discipline-specific landing pages describing data published through GFZ Data Services. Example for the publication service for global gravitational models of ICGEM, the International Centre for Global Earth Models [42], (**left**). Project-specific landing page for hyperspectral airborne data within the German Environmental Mapping and Analysis Program (EnMAP) [43], (**right**). See Figure 3 for the standard GFZ layout. The map data are © OpenStreetMap contributors and published under a CC-BY-SA license.

6. Licenses and Performance of the Software Stack

Depending on the targeted audience of an intellectual work it can be published under a license that sets limits to reuse or does not set any limit to future applications. Existing licenses for software can determine the licensing terms for derived software. Therefore, particular care needs to be taken to respect the licenses applied by each of the original authors when combining different software components. In our case the eSciDoc middleware uses the "Common Development and Distribution License" in version 1.0, which is also used by the PubMan application. The DOIDB and panMetaDocs are published under "Apache License" in version 2.0. The metadata editor uses the "GNU General Public License" in version 3.0 that originates from the Sencha EXTJS framework and is in our case the most restrictive license. All software components—eSciDoc, DOIDB, panMetaDocs, PubMan, and the metadata editor are combined using web services, which allows us to separate different license domains since none of the licenses disperses through web APIs.

To estimate the performance of the whole software stack we have to look at the DOIDB and the eSciDoc middleware and their applications separately. The DOIDB acts as a proxy to the DataCite infrastructure and the time needed to process registration calls reflects the response time of the DataCite

infrastructure. In our case DOIDB registrations take in the order of 1.2 to 1.5 s and there is no significant difference between direct registration at DataCite and registration through our DOIDB. The search and dissemination components operate locally without communication to external infrastructure and the response time is typically in the order of fractions of a second.

The eSciDoc middleware is the central storage for several applications that use the eSciDoc REST API to store data and metadata. There is no external infrastructure involved to store data or metadata. In case of performance problems, the migration of applications that use eSciDoc as data backend to a different and faster host is very easy. However, the eSciDoc middleware cannot be replicated and has to be set up in a reasonably fast environment. Our collection currently consists of approximately 28,000 eSciDoc items and read operations to download the metadata of a single item typically take around one second. Our software stack allows the storage of files up to 100 MB, which is a limit set by the PHP runtime environment of panMetaDocs. The largest file we store is 2.8 GB in size and has been uploaded through a Java program. For very large datasets we store only metadata in the eSciDoc middleware and make use of eSciDocs ability to point to external servers for binary content or we show a contact form and use alternative transfer protocols.

7. Discussion, Lessons Learned, and Future Work

The publishing infrastructure presented in this article consists of a central storage facility for data and metadata (eSciDoc), applications to upload data and metadata into the storage facility (PubMan, panMetaDocs), and an application to log DOI registrations and associated metadata (DOIDB). This publishing infrastructure was built in a modular fashion to allow single components to be exchanged with no or only minor effects on the operational status of the system. However, eSciDoc as the central storage plays a pivotal role in the system and all applications using it for data storage must be compatible with the eSciDoc API. This is important when changes are made to the eSciDoc API in the course of version updates.

To minimize maintenance costs, we sought to avoid bespoke software developments and the risk of vendor lock-in, but to reuse existing free and open source components developed by the digital curation community. The resulting publishing system consists of pre-existing software components that already fit most of our needs and needed little adaptation.

There is a certain degree of redundancy in the way metadata are stored in each of the components, that is in eSciDoc, in DOIDB, and respective panMetaDocs instances. However, each metadata store in the software components serves a specific purpose: eSciDoc stores data and metadata related to items, the DOIDB stores metadata of GFZ DOIs, and panMetaDocs stores a copy of the metadata of the respective instance to bypass the global access rights management of eSciDoc in favour of project based access to metadata of unpublished data.

The delegation of the user authentication into the GFZ LDAP saved us from administering user accounts by reusing information that is already provided by the identity management of GFZ IT Services. The downside of using LDAP is that it ties the user management of the eSciDoc infrastructure, and all applications connected to eSciDoc, to users at GFZ while making it difficult to include external users. In some cases, external access to data is required, be it in project collaborations or in a peer-review process involving data. To allow reviewers to access data, a temporary link can be created to allow access to unpublished data. However, the problem persists and it is cumbersome to administer external user accounts in eSciDoc when LDAP is the main authentication mechanism. At time of implementation of eSciDoc at GFZ, user account data were only available through LDAP. For new installations, we recommend to use technologies for federated authentication, like OpenID and Shibboleth [23,24].

PHP was chosen for eSciDoc client applications because it seemed to be easier to use by researchers in projects, compared to programming languages like Java. Today we would probably recommend Python because of its low entry barrier and rich catalogue of scientific code libraries. Handling large streams of data in PHP is a rather exotic use case for this scripting language. PHP is designed for fast delivery of web sites and large file sizes turned out to be problematic. While changing memory

management parameters of the PHP runtime helps to solve this problem, the PHP runtime environment limits the maximum duration of uploads. This turned out to be problematic because the duration needed to upload data is difficult to estimate in advance. A possibility to handle large data is to store the XML metadata inside eSciDoc and to store a reference (URL) to the binary data on an external server. This approach avoids large uploads to eSciDoc but also bypasses eSciDocs access control to binary data.

While community specific landing pages are a requirement in federated projects, maintaining their web presence over long time periods also requires additional resources for maintenance of the underlying infrastructure. This has to be monitored carefully, although we believe that XML Stylesheet Transformation (XSLT) can minimize maintenance requirements by providing a mechanism independent of hardware platform and programming language to generate web pages for the presentation of data.

Currently, panMetaDocs' metadata editor provides assistance for the entry of spatial coverages. The form could also provide support to fill in correct identifiers by an interactive search for related datasets, related articles, and authors through the APIs of DataCite, Crossref, and ORCID. This would help researchers to link their datasets to related data, manuscripts, and authors. Furthermore, there is a demand for research domain specific extensions or modifications of the metadata editor. It is possible to open the metadata for linked data applications and enable this way machine to machine communication. There is a mapping from the DataCite Metadata Schema to the Resource Description Framework (RDF) and such mappings are also available for ISO19139 metadata [44,45]. We make use of terms from the GEMET multi-lingual thesaurus used for INSPIRE metadata [27]. In addition, we apply science keywords from NASA GCMD DIF [28] to the metadata. Both vocabulary lists are very popular and make it easier to find related data.

8. Conclusions

Based on the concept of "Domains of Responsibility" [15], we established a data curation and publication infrastructure that is scalable and adaptable to the ever-changing requirements of research projects in a national research centre. The modular nature of the system, and its service oriented architecture, allows us to maintain large parts of the system unchanged, while adding new components and decommissioning obsolete components as needed. Software components may be exchanged as long as its replacement and our client components share a common API. Using a contents agnostic middleware to manage data storage allows us to process any file type described by any metadata schema, including by multiple concurrent schemas, coming in multiple representation forms, or being stored outside the system. All software components are based on free and open source software, avoiding bespoke developments where possible.

Providing a common API allows projects to develop their own data systems, independently of the enterprise data management system. For transfer of data into the "Persistent Domain", independent systems need to be able to interface with the common API. Stored data are safe and curated by the library, even if the maintenance of the project specific software may not continue one day. From an organisational perspective, this separation of concerns into "Domains of Responsibility" [14] has the advantage that it becomes easier to allocate resources for system development and maintenance, and determine sources of funding these resources. While it may seem difficult to maintain software from completed projects, this may no longer be necessary in the presented scenario. All data and metadata should then have been transferred into the "Persistent Domain" operated by the "memory institution" where they will remain curated and accessible.

The pattern presented in this paper is in line with current concepts of research data curation and long-term digital preservation. Its modular and services oriented architecture can be adapted to the changing needs of data curation in research projects, while at the same time maintaining a stable core data storage and access infrastructure that develop at a much slower pace. This pattern may therefore

serve as a blueprint to other institutions planning to implement an enterprise-wide system for research data curation.

Author Contributions: Jens Klump and Roland Bertelmann designed and established the central repository for GFZ Data Services. In addition, they coordinated activities for the implementation of dataset DOIs at GFZ. Kirsten Elger was involved in the design of the metadata editor and the presentation pages and is currently responsible for the publication of research data at GFZ. Damian Ulbricht was involved in the software development of panMetaDocs, the DOIDB, the Javascript metadata editor, and the XSLT stylesheets for presentation of datasets. The drafting of the manuscript was a joint effort of all authors.

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

References

1. Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. Available online: http://openaccess.mpg.de/Berlin-Declaration (accessed on 24 February 2016).
2. G8 Science Ministers. G8 Science Ministers Statement. Available online: https://www.gov.uk/government/news/g8-science-ministers-statement (accessed on 10 June 2015).
3. European Commission. EU Implementation of the G8 Open Data Charter. Available online: http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=3489 (accessed on 10 June 2015).
4. Federal Ministry for Economic Affairs and Energy, Federal Ministry of the Interior and Federal Ministry of Transport and Digital Infrastructure. Digital Agenda 2014–2017 of the Federal Government of Germany. Available online: http://www.digitale-agenda.de/Content/DE/_Anlagen/2014/08/2014-08-20-digitale-agenda-engl.pdf?__blob=publicationFile&v=6 (accessed on 10 June 2015).
5. Organisation for Economic Co-operation and Development (OECD). Declaration on Access to Research Data from Public Funding. Available online: http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=157 (accessed on 26 November 2015).
6. European Parliament, Council of the European Union. Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). Available online: http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32007L0002 (accessed on 28 November 2015).
7. Coalition on Publishing Data in the Earth and Space Sciences. Statement of Commitment from Earth and Space Science Publishers and Data Facilities. Available online: http://www.copdess.org/statement-of-commitment/ (accessed on 5 July 2015).
8. Klump, J. Geowissenschaften. In *Langzeitarchivierung von Forschungsdaten—Eine Bestandsaufnahme*; Neuroth, H., Strathmann, S., Oßwald, A., Scheffel, R., Klump, J., Ludwig, J., Eds.; Verlag Werner Hülsbusch: Boizenburg, Germany, 2012; pp. 179–194.
9. GEOFON Program, GFZ Potsdam. Available online: http://geofon.gfz-potsdam.de (accessed on 30 November 2015).
10. International GNSS Service IGS (Formerly the International GPS Service). Available online: https://igscb.jpl.nasa.gov (accessed on 30 November 2015).
11. International Real-time Magnetic Observatory Network. Available online: http://www.intermagnet.org (accessed on 30 November 2015).
12. Kratz, J.; Strasser, C. *Data Publication Consensus and Controversies*; F1000Research: London, UK, 2014.
13. GFZ Publication Database. Available online: http://gfzpublic.gfz-potsdam.de (accessed on 30 November 2015).
14. Klump, J.; Ulbricht, D.; Conze, R. Curating the web's deep past—Migration strategies for the German Continental Deep Drilling Program web content. *GeoRes. J.* **2015**, *6*, 98–105. [CrossRef]

15. Treloar, A.; Groenewegen, D.; Harboe-Ree, C. The data curation continuum—Managing data objects in institutional repositories. *D-Lib Mag.* **2007**, *13*. [CrossRef]

16. Devarakonda, R.; Palanisamy, G.; Green, J.; Wilson, B. Data sharing and retrieval using OAI-PMH. *Earth Sci. Inf.* **2011**, *4*, 1–5. [CrossRef]

17. Fielding, R.T. Architectural Styles and the Design of Network-based Software Architectures. Ph.D. Thesis, University of California, Irvine, CA, USA, 2000.

18. Huber, R. panMetaWorks—PangaWiki. Available online: http://wiki.pangaea.de/wiki/Panmetaworks (accessed on 9 July 2015).

19. Ulbricht, D.; Klump, J. panMetaDocs—A tool for collecting and managing digital objects in a scientific research environment. In Proceedings of the European Geosciences Union General Assembly, Vienna, Austria, 3–8 April 2011.

20. Ulbricht, D.; Klump, J.; Bertelmann, R. Publishing datasets with eSciDoc and panMetaDocs. In Proceedings of the European Geosciences Union General Assembly, Vienna, Austria, 22–27 April 2012.

21. Razum, M.; Schwichtenberg, F.; Wagner, S.; Hoppe, M. eSciDoc infrastructure: A Fedora-based e-research framework. In Proceedings of the European Conference on Digital Libraries, Corfu, Greece, 27 September–2 October 2009; pp. 227–238.

22. The Internet Society. Lightweight Directory Access Protocol (LDAP): Technical Specification Road Map. Available online: https://tools.ietf.org/html/rfc4510 (accessed on 9 July 2015).

23. The OpenID Foundation. Specifications & Developer information. Available online: http://openid.net/developers/specs/ (accessed on 9 July 2015).

24. The Shibboleth Consortium. What's Shibboleth. Available online: http://shibboleth.net/about/ (accessed on 9 July 2015).

25. panMetaDocs Source Code Repository. Available online: http://panmetadocs.sf.net (accessed on 29 November 2015).

26. Starr, J.; Gastl, A. IsCitedBy: A metadata scheme for DataCite. *D-Lib Mag.* **2007**, *17*. [CrossRef]

27. Drafting Team Metadata and European Commission Joint Research Centre. *INSPIRE Metadata Implementing Rules: Technical Guidelines Based on EN ISO 19115 and EN ISO 19119*; European Commission Joint Research Centre: Brussels, Belgium, 2010. Available online: http://inspire.ec.europa.eu/documents/Metadata/INSPIRE_MD_IR_and_ISO_v1_2_20100616.pdf (accessed on 20 February 2016).

28. National Aeronautics and Space Administration—Global Change Master Directory. Directory Interchange Format (DIF) Writer's Guide. Available online: http://gcmd.nasa.gov/add/difguide/ (accessed on 28 November 2015).

29. Klump, J.; Huber, R.; Diepenbroek, M. DOI for geoscience data—How early practices shape present perceptions. *Earth Sci. Inf.* **2016**, *9*, 123–136. [CrossRef]

30. Klump, J.; Bertelmann, R.; Brase, J.; Diepenbroek, M.; Grobe, H.; Höck, H.; Lautenschlager, M.; Schindler, U.; Sens, I.; Wächter, J. Data publication in the Open Access Initiative. *Data Sci. J.* **2006**, *5*, 79–83. [CrossRef]

31. Klump, J.; Ulbricht, D. Re-using the DataCite metadata store as DOI registration proxy and IGSN registry. In Proceedings of the American Geophysical Union, San Francisco, CA, USA, 3–7 December 2012.

32. DataCite Metadata Working Group. *DataCite Metadata Schema for the Publication and Citation of Research Data*; version 3.1; DataCite: London, UK, 2015; doi:10.5438/0010. [CrossRef]

33. DOIDB Store GitHub Repository. Available online: https://github.com/ulbricht/mds/tree/doidb (accessed on 29 November 2015).

34. DOIDB Search GitHub Repository. Available online: https://github.com/ulbricht/search/tree/doidb (accessed on 29 November 2015).

35. DOIDB OAI-PMH GitHub Repository. Available online: https://github.com/ulbricht/oaip/tree/doidb (accessed on 29 November 2015).

36. National Information Standards Organization. *Understanding Metadata*; NISO Press: Bethesda, MD, USA, 2004.

37. Library of Congress. Recommended Formats Statement. Available online: http://www.loc.gov/preservation/resources/rfs/ (accessed on 24 February 2016).

38. Metadata Editor for panMetaDocs GitHub Repository. Available online: https://github.com/ulbricht/pmdmeta (accessed on 29 November 2015).

39. Barthelmes, F.; Köhler, W. International Centre for Global Earth Models (ICGEM). In *Journal of Geodesy: The Geodesist's Handbook 2012*; Drewes, H., Ed.; Springer: Berlin, Germany, 2012; pp. 932–934.

40. Guanter, L.; Kaufmann, H.; Segl, K.; Foerster, S.; Rogaß, C.; Chabrillat, S.; Küster, T.; Hollstein, A.; Rossner, G.; Chlebek, C.; *et al*. The EnMAP spaceborne imaging spectroscopy mission for earth observation. *Remote Sens.* **2015**, *7*, 8830–8857. [CrossRef]

41. Lorenz, H.; Rosberg, J.-E.; Juhlin, C.; Bjelm, L.; Almqvist, B.; Berthet, T.; Conze, R.; Gee, D.G.; Klonowska, I.; Pascal, C.; *et al*. *COSC-1 Operational Report—Operational Data Sets*; GFZ Data Services: Potsdam, Germany, 2015. Available online: http://doi.org/10.1594/GFZ.SDDB.ICDP.5054.2015 (accessed on 20 February 2016).

42. Förste, C.; Bruinsma, S.; Abrikosov, O.; Lemoine, J.-M.; Marty, J.C.; Flechtner, F.; Balmino, G.; Barthelmes, F.; Biancale, R. *EIGEN-6C4 The Latest Combined Global Gravity Field Model Including GOCE Data up to Degree and Order 2190 of GFZ Potsdam and GRGS Toulouse*; GFZ Data Services: Potsdam, Germany, 2015. Available online: http://doi.org/10.5880/icgem.2015.1 (accessed on 20 February 2016).

43. Neumann, C.; Weiss, G.; Itzerott, S. *Döberitzer Heide 2008/2009—An EnMAP Preparatory Flight Campaign*; GFZ Data Services: Potsdam, Germany, 2015. Available online: http://doi.org/10.5880/enmap.2015.001 (accessed on 20 February 2016).

44. DataCite Ontology GitHub Repository. Available online: http://github.com/datacite/ontology (accessed on 17 January 2016).

45. OWL Representation of ISO 19115 (Geographic Information-Metadata). Available online: http://def.seegrid. csiro.au/isotc211/iso19115/2003/metadata (accessed on 24 February 2016).