Earth Science Informatics (2020) 13:641–654
https://doi.org/10.1007/s12145-020-00441-0

RESEARCH ARTICLE

Managing collaborative research data for integrated, interdisciplinary environmental research

M. Finkel[1] · A. Baur[2] · T.K.D. Weber[3] · K. Osenbrück[1] · H. Rügner[1] · C. Leven[1] · M. Schwientek[1] · J. Schlögl[1] · U. Hahn[2] · T. Streck[3] · O.A. Cirpka[1] · T. Walter[2] · P. Grathwohl[1]

Received: 23 September 2019 / Accepted: 1 January 2020 / Published online: 16 January 2020

Abstract

The consistent management of research data is crucial for the success of long-term and large-scale collaborative research. Research data management is the basis for efficiency, continuity, and quality of the research, as well as for maximum impact and outreach, including the long-term publication of data and their accessibility. Both funding agencies and publishers increasingly require this long term and open access to research data. Joint environmental studies typically take place in a fragmented research landscape of diverse disciplines; researchers involved typically show a variety of attitudes towards and previous experiences with common data policies, and the extensive variety of data types in interdisciplinary research poses particular challenges for collaborative data management. In this paper, we present organizational measures, data and metadata management concepts, and technical solutions to form a flexible research data management framework that allows for efficiently sharing the full range of data and metadata among all researchers of the project, and smooth publishing of selected data and data streams to publicly accessible sites. The concept is built upon data type-specific and hierarchical metadata using a common taxonomy agreed upon by all researchers of the project. The framework's concept has been developed along the needs and demands of the scientists involved, and aims to minimize their effort in data management, which we illustrate from the researchers' perspective describing their typical workflow from the generation and preparation of data and metadata to the long-term preservation of data including their metadata.

Keywords Research data management · Interdisciplinary environmental research · Metadata · Taxonomy

Introduction

There is broad worldwide consensus that publicly funded science should be open and research knowledge should be shared (e.g., Berlin Declaration 2003, Nosek et al. 2015). This is reflected by a large number of initiatives and networks, on institutional, national, and international level, such as (i) the European Open Science Cloud (EC 2017), (ii) the Open Knowledge Foundation (https://okfn.org), which put forward the Open Definition to exactly define the meaning of "open" (https://opendefinition.org) and the Open Data Handbook (http://opendatahandbook.org), (iii) the Center for Open Science (https://cos.io), which provides guidance and tools to foster the distribution of knowledge and data and are leading and supporting projects that contribute to open data, (iv) the expert group on 'Science 2.0/Open Science' established by the European University Association (https://eua.eu), (v) the Open Scholarship Initiative (OSI) in partnership with UNESCO (http://osiglobal.org), (vi) the 'cOAlition S' of European research funding organizations (https://www.coalition-s.org), and (vii) the Coalition for Publishing Data in the Earth and Space Sciences (COPDESS, https://copdess.org) including the American Geophysical Union (AGU) and

Communicated by: H. Babaie

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12145-020-00441-0) contains supplementary material, which is available to authorized users.

✉ M. Finkel
michael.finkel@uni-tuebingen.de

[1] Center for Applied Geoscience, University of Tübingen, Hölderlinstrasse 12, 72074 Tübingen, Germany

[2] Central Data Administration, Department of Informatics, University of Tübingen, Wächterstraße 76, 72074 Tübingen, Germany

[3] Institute of Soil Science and Land Evaluation, University of Hohenheim, 70593 Stuttgart, Germany

the European Geoscience Union (EGU) who require that data related to publications of these associations must be accessible via open data repositories. Among the key objectives of the Open Science movement are the provision of Open Access to research results and Open Data (which, according to the Open Definition, implies the free use, re-use, and redistribution of data by anyone – subject only, at most, to the requirement to attribute and share alike). Obviously, these objectives can be achieved only if research data management and stewardship meet certain criteria with respect to the collection, annotation, long-term preservation and archiving of data. Corresponding guiding principles, defined by Wilkinson et al. (Wilkinson et al. 2016, 2018), are commonly referred to as the FAIR Data Principles with Findability, Accessibility, Interoperability, and Reusability being the four fundamental dimensions.

In recent years, a large number of research-data-management platforms and solutions have been developed. The registry of research data repositories (https://www.re3data.org/) as of November 2019 lists 269 repositories for 'environmental data'. These research-data-management solutions differ with respect to technical architecture, metadata, user and programming interfaces, scope, coverage, and costs (e.g., Glatard et al. 2017; Amorim et al. 2017). Research data management is a complex issue involving multiple activities carried out by various actors addressing a range of drivers and influenced by a large set of factors (Pinfield et al. 2014). An ongoing debate addresses the roles and responsibilities of these actors, and how they are best allocated for the diverse tasks and activities of research data management (White 2014; Latham 2017; Chunpir 2018; Rodrigues et al. 2019).

The proper – i.e. efficient and consistent – management of research data is not only required to meet the overarching goal of open science and its expected benefits for promoting research worldwide, it also plays a crucial role for the success of individual research activities themselves. This is particularly true for collaborative research that aims to achieve an integrated view of the research subjects and typically involves many researchers from different disciplines. For large and long-term research consortia, and for research that deals with a variety of data in terms of type, size, quality, etc., the employment of an efficient as much as effective infrastructure for the management of research data is mandatory to assure successful high-quality research (e.g., Geosling et al. 2015; Specht et al. 2015; Wang et al. 2016; Curdt 2016, and 2019; Grunzke et al. 2019).

A key challenge of handling data in interdisciplinary environmental research projects is the variety of data types, including stationary geodata (e.g., topography, geology, soil maps), streaming data (e.g., from sensors with data loggers), data from external sources (such as weather services), geophysical surveying data, data related to individual soil, rock, biomass or water samples taken in the field, data from lab experiments, among others. Often the annotation of metadata (such as the origin of a sample, type and conditions of experiments, applied measurement technique and device used, calibration method) is equally important as the data values themselves (e.g., Hsu et al. 2015). Given the variety of activities, two data sets taken by two research teams may look completely different, because they follow different, even though internally coherent reasoning (e.g., White 2014).

A second key challenge typical for interdisciplinary environmental research is the variation in attitudes of the individual researchers involved towards a common data policy. In research fields, in which large data streams of well-defined origins of a few data types are collected, and which can only be interpreted in a meaningful way when data streams of different agencies are combined (such as in seismology or meteorology), common data formats have been unanimously accepted, as a prerequisite for analyzing data, operational forecasting, and conducting research with the data. In more fragmented research fields, such as in large-scale interdisciplinary environmental research projects, it is much more difficult to define common grounds for data management. While it is clear that a holistic analysis of the data collected in these kinds of projects is only possible if all data and metadata are accessible on a common data platform, not every individual researcher providing data to the platform is involved in such holistic analyses, or needs access to data from many other researchers within the consortium. This hampers the intrinsic motivation to contribute to a common data management scheme (Dehnhard et al. 2013; Tenopir et al. 2015; Fecher et al. 2015; Kratz and Strasser 2015). We claim that this situation is fairly typical for large parts of the environmental research landscape.

In this paper, we propose a data management framework that meets the specific challenges of interdisciplinary environmental research outlined above. We present several novel features in order to meet the requirements of collaborative research data management in environmental research as defined by the researchers' needs and the goals of integrated research. The framework is being developed within the collaborative research center (CRC) CAMPOS 'Catchments as Reactors: Metabolism of Pollutants on the Landscape Scale' (https://www.campos.uni-tuebingen.de), funded by the German Research Foundation (DFG), which started in January 2017 and may run in three phases up to 12 years. Major parts of the framework have already been implemented and in operation, supporting ongoing research work. We believe that the flexibility of our approach enables its concepts to be transferred to other projects of similar scientific orientation and may give inspiration to or serve as a model for research data management in future projects.

Measures and methods

Organizational measures

The constitution of an efficient organizational framework is a prerequisite for the development, implementation, and operation of efficient research data management within a project. Towards this end, we have taken the following organizational measures within CAMPOS:

(i) *Focus on data management from the beginning*: We started to build awareness of data management requirements and goals within the consortium from the beginning of the project proposal activities. Within the proposal we emphasized the importance of the data management infrastructure for the success of the proposed project and made clear that data management means more than just a number of service tasks. This includes the commitment to concertedly conceptualize and develop procedures and tools to implement the integrated data management of the project. The obligation of all members of the CAMPOS consortium to document research data is stated in the bylaws of the consortium agreement.

(ii) *Data Management as research and service*: We structured the required work into a research part with a focus on the development of data management concepts and the infrastructure, and a service part, which deals with the analyses and consideration of researchers' demands as well as all tasks required to coordinate the implementation and operation of the data management infrastructure within the consortium, including communication and training of the researchers. This distribution of research and service tasks, which are very different in content and form, is a key requirement for an effective development and implementation of the data management within the CRC as well as for the appropriate distribution of personnel resources. A prerequisite is the allocation of sufficient resources from the overall budget.

(iii) *Appropriate organizational entities*: Organizational entities have been constituted on three levels to efficiently coordinate the development and implementation of the data management within the CRC: (a) top level decisions are made in the 'Executive Board' concerning, amongst others, strategic issues, prioritization of activities, and type of data licensing; (b) 'project data managers' (PDMs) are responsible for the organization of work required to develop and implement the data management within the individual projects of the CRC and meet regularly to update, determine and coordinate working tasks, which are to be implemented within individual projects; (c) 'data teams' have been constituted to discuss the needs and options to manage specific types of data, including details on metadata definitions, controlled vocabularies, to name a few; (d) responsible persons in the data infrastructure and data service project and the project data managers form the 'Research-Data-Management Core Group', the members of which have extended rights and responsibilities (such as the maintenance and further development of taxonomy of terms used in data descriptions) in research data management (Fig. 1).

(iv) *Maximum communication*: The desired involvement of the entire consortium in data management activities in order to efficiently distribute the variety of tasks among researchers, technical and service personnel requires a maximum in communication within the consortium. In addition to regular meetings of the 'Executive Board', project data managers, and data teams, which form the core body of communicating data management issues, we have taken further measures to promote communication about data management: (a) data management is periodically on the agenda of a weekly research seminar; (b) specific workshops are continuously organized and held to set a focus on particular topics (such as tools supporting metadata generation, handling of particular types of data, e.g., from non-target chemical analysis, structural/functional analysis of organisms ("omics"), or depth-oriented field investigation); (c) electronic manuals on the use of the data management infrastructure have been issued and are regularly updated.

The data management concept: Core elements and structure

The research in the CAMPOS project focuses on diffuse pollution of soils, surface waters, and groundwater by a multitude of anthropogenic contaminants and their turnover at landscape scale. It consists of eight collaborative projects that differ considerably with respect to their research goals, the specialization of their personnel, the way data is typically dealt with, existing workflows, and the type of data that is being produced. The spectrum of research data ranges from concentration measurements of specific compounds (so called target analysis), quantification of high-resolution molecular fragments in water samples (so called non-target analysis), outcomes of toxicity tests performed on surface-water samples, hydrological, geological, and hydrogeological data obtained as monitoring data or during field tests, to genomic and metabolomics data from molecular-biological analysis of samples. The produced data sets differ in terms of size, dimension, structure, format, temporal frequency, and origin, among others. The observed data are used to inform and calibrate numerical models that simulate water fluxes and reactive solute transport within the catchment and its compartments.

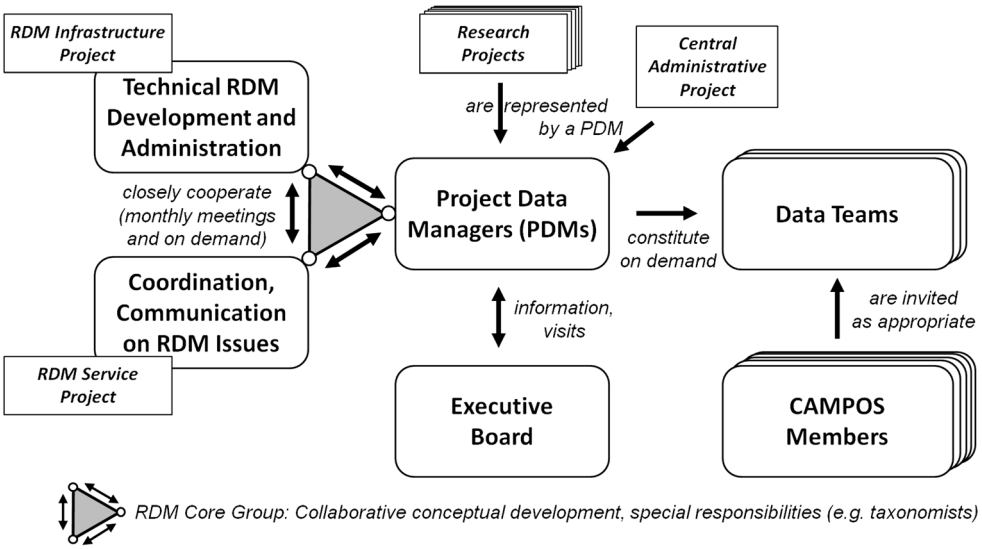

Fig. 1 Scheme of organizational entities for the conceptualization, development and implementation of data management within CAMPOS

Numerical model results, obtained in both deterministic and stochastic model runs, contribute large data volumes to be documented, stored, and maintained. The research results also comprise software codes, developed for data analysis and process modeling.

To answer the challenges given by the diverse spectrum of disciplines involved in the CAMPOS project and the multitude of existing workflows, as well as to serve the researchers' needs, we defined a general data management framework (Fig. 2) based on five core conceptual elements described below.

Ongoing research data vs. publicly available data

We distinguish between a project-internal area, in which all research data is managed, and a public area. The latter is planned to consist of a public web portal run by CAMPOS and the long-term Research Data Archive administrated by the University of Tübingen. The CAMPOS-internal area provides a 'safe environment' for all CAMPOS research data. Access is limited to registered members of the consortium, with clear rules with respect to intellectual property rights, ownership, and use of data, which are stipulated in the data management plan. The mutual management of data guarantees early provision of data for integrated interpretations and analyses. The data scope covers all data relevant for the integrated research within CAMPOS: newly generated data, data from previous projects, and data from external sources (e.g., weather data from Germany's National Meteorological Service, DWD, and geological- and hydrogeological data from state agencies). Two gateways allow making data publicly available:

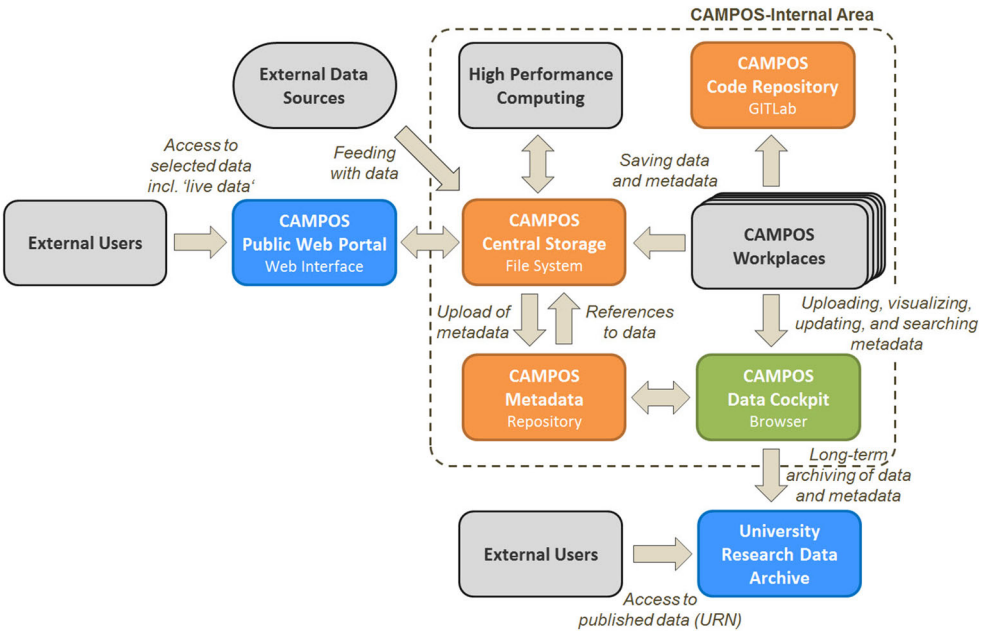

Fig. 2 Structure of the CAMPOS data management framework

the direct provision and display of selected data, for example, climate data and soil status data for farmers, will be done via the CAMPOS Public Web Portal that directly accesses the data and its metadata from the database and file system of the CAMPOS-internal area. The publication and archiving of data is accomplished via the university's long-term research data archive, the Research Data Portal FDAT (Kaminski and Brandt (2018) and https://fdat.escience.uni-tuebingen.de/portal/). The ingest process for the archiving of data, from registration via metadata annotation, data package bundling, and verification to publication is done to a large extent automatically on the basis of the data and metadata stored in the internal area (see also section 2.3).

Separation of data and metadata

Data and metadata are stored separately. This separation, with data being stored in a file system, and metadata being managed in a database increases the flexibility and efficiency of data management. Furthermore, it is an important prerequisite to meet several essential requirements of model-based integrated environmental research, such as (i) the direct access to data for modelling and data processing codes, (ii) fast search for data – in the CAMPOS Data Cockpit – using related metadata organized in a slim and fast database, (iii) convenient storage of huge data pieces (such as from non-target and genomic analyses), (iv) convenient and reliable data updating without any access or even changes to the metadata repository.

More streamlined approaches are possible and recommendable for data with a rather specific profile with respect to these criteria (e.g., like the hydrologic information system (HIS) launched by the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI), which is based on a standard database scheme for storing point observations with a relational database (e.g., Horsburgh et al. 2008, 2015; Sadler et al. 2016)). However, we believe that these approaches are of limited use for the consistent management of heterogeneous data in multidisciplinary research.

While well-defined databases have known strengths and benefits in terms of performance scalability and search capabilities, their suitability is limited when flexible solutions are required to store data from a continuously evolving research environment with data-formats and -types, which are not known a priori but increase in number over time. Object-based data storage solutions, which have gained popularity as an efficient and performant alternative in specific cases (e.g., Blaylock et al. 2017), do not provide the necessary flexibility.

Hierarchically and flexibly structured data type-specific metadata

In order to avoid redundancy and inconsistencies in metadata we follow a hierarchical concept for metadata creation. As examples, we distinguish between metadata associated to a measurement location and metadata associated to individual samples/probes at these locations, or – in very much the same sense – metadata of a lab experiment using material of this sample (Fig. 3). This concept offers large flexibility and efficiency because metadata can be defined in a datatype-specific way. To follow above example, we separate (a) metadata needed to describe the installation of a groundwater well at a certain location (including coordinates, drilling company, depth, etc.) from (b) metadata associated to periodic water quality analyses (sampling times, preservation of samples, analytical procedures in the lab), from (c) metadata describing other monitoring measurements (such as hydraulic-head measurements) or (d) field experiments (such as well tests, requiring information about pumping rates and duration) from (e) metadata describing laboratory experiments performed on soil samples taken upon well installation. This concept of splitting metadata into pieces (that are logically linked via identifiers) allows accounting for and tying in with existing procedures, protocols, and documentation standards, which vary among the different activities and data types. Along these lines, we have designed metadata templates for individual types of data or activities. For most individual researchers this means that their contribution to data management can be restricted to their particular research data and description of procedures, minimizing their efforts by avoiding any additional, unnecessary expense.

Two-step metadata creation

The creation of metadata is further streamlined by a two-step process of metadata creation. In a first step, CAMPOS members create metadata files (formatted as OASIS-OpenDocument or Office Open XML) using a standard spreadsheet-calculation software, applying a common general metadata structure (for further details see section 3.3). These metadata are transferred from the workplace of the respective researcher or technician to the file system of the CAMPOS Central Storage – together with the corresponding data. In a second step the metadata is uploaded into the database. This upload process includes an automated validation of the metadata (see also section 3.4).

Taxonomy of terms used in metadata

To ensure the creation of consistent metadata that can be accurately and quickly searched and retrieved, we use controlled vocabularies in descriptive metadata fields (e.g., Hedden 2010). Both flat and hierarchical control schemes are used to define the taxonomy of accepted terms. This includes the definition of synonyms (also referred to as non-preferred aliases, see Hedden 2010) to enable a flexible search and to overcome ontological and semantic heterogeneity when data is

synthesized with other repositories, as discussed, e.g., in Piasecki and Beran (2009) and in Horsburgh et al. (2014). All terms are in English. Controlled vocabularies are defined during the creation of data-type specific metadata templates if appropriate (see section 3.3). Existing vocabularies may be adopted, for example, if data from external sources is imported. A central taxonomy service, maintained by the administrator of the CAMPOS Data Cockpit, offers convenient access to the vocabularies for all researchers and interactive editing capabilities for the taxonomists group (RDM Core Group, see Fig. 1) to continuously update the vocabularies upon users' demand.

Technical infrastructure

The CAMPOS framework for management of research data consists of three functional environments: (a) the *CAMPOS-Internal Area* forming the private working environment of the CAMPOS researchers for all data-related tasks and issues, (b) the *Research Data Archive FDAT* of the University of Tübingen to preserve and publish data for long-term storage and use, and (c) the *CAMPOS Public Web Portal* to provide public access to selected data.

CAMPOS internal area

The CAMPOS Internal Area consists of three main components:

1. The *CAMPOS Central Storage*, hosted at the central computing facilities of the University of Tübingen, is a file system distributed as a SMB shared network drive with an effective capacity of 100 TB that stores all relevant internal research data of the CAMPOS project. It implements a predefined filesystem structure and serves as the starting point for data ingest into the Data Cockpit. Differential backups are created every night and retained for 30 days (including all relevant media like database dumps and operating system snapshots). Remote access is accomplished via WAN connections using a tunnel (VPN) service. Access control and file system permissions on the SMB network drive are set up using Lightweight Directory Access Protocol (LDAP) group policies and Windows access control lists allowing a fine-grained definition of user privileges in a convenient way. Please note that, in addition to the central storage, relational (PostgreSQL) and NoSql (MongoDB) databases are in use where appropriate. Data like variable structured continuous data streams delivered by measurement devices go into a NoSql store. Well-structured existing data that will not undergo any further changes is managed within a relational database.
2. The *CAMPOS Metadata Repository* is implemented as a relational database and holds all registered metadata sets including references to the actual data stored in the file system. It serves as the basis for data search.
3. The *CAMPOS Data Cockpit* is a web interface implemented as a Ruby on Rails (https://rubyonrails.org/) plugin for the web application Redmine (http://www.redmine.org) taking advantage of existing functionality and adding missing workflows where needed. The Data Cockpit provides access to data and metadata for all CAMPOS internal users (see also section 3.4 below). The CAMPOS Data Cockpit interfaces the Public Web Portal and the Research Data Archive FDAT. An automated procedure bundling data and metadata to build specific ingest packages facilitates the data publication and archiving process. Ingest packages include – in addition to the detailed data type-specific metadata – a fixed set of FDAT-specific metadata required for ingest and archiving of data in FDAT, as well as for search functionalities and display in the web front end of FDAT. This FDAT-specific metadata is formatted as XML file for transfer making use of Metadata Encoding & Transmission Standard (METS), Encoded Archival Description (EAD) and Preservation Metadata Implementation Strategies (PREMIS).

All services of the internal area are accessible for CAMPOS members only. The user and account management system for authentication of all CAMPOS infrastructure services includes user information for all members distinguishing between staff of the hosting institution (University of Tübingen) and members affiliated with partner institutions (University of Hohenheim, University of Stuttgart, Technical University Munich, Helmholtz Center for Environmental Research Leipzig). Parts of the existing infrastructure of the central computing facilities of the University of Tübingen (a filtered LDAP facade) are combined with a new LDAP instance holding all members of partner institutions.

Research data archive FDAT

The Research Data Archive FDAT consists of the open source repository software Fedora Commons (https://duraspace.org/fedora) interfacing a web front end that provides public access to the archived resources. FDAT accepts ingest packages in the form of compressed file bundles containing the resources to be archived and an XML file (Metadata Encoding & Transmission Standard (METS) encapsulating an Encoded Archival Description (EAD) and the PREMIS Data Dictionary, see https://www.loc.gov/standards/premis/). A persistent, location-independent resource identifier, a uniform resource name (URN) is assigned to each data resource. A versioning of the data in FDAT has not been implemented

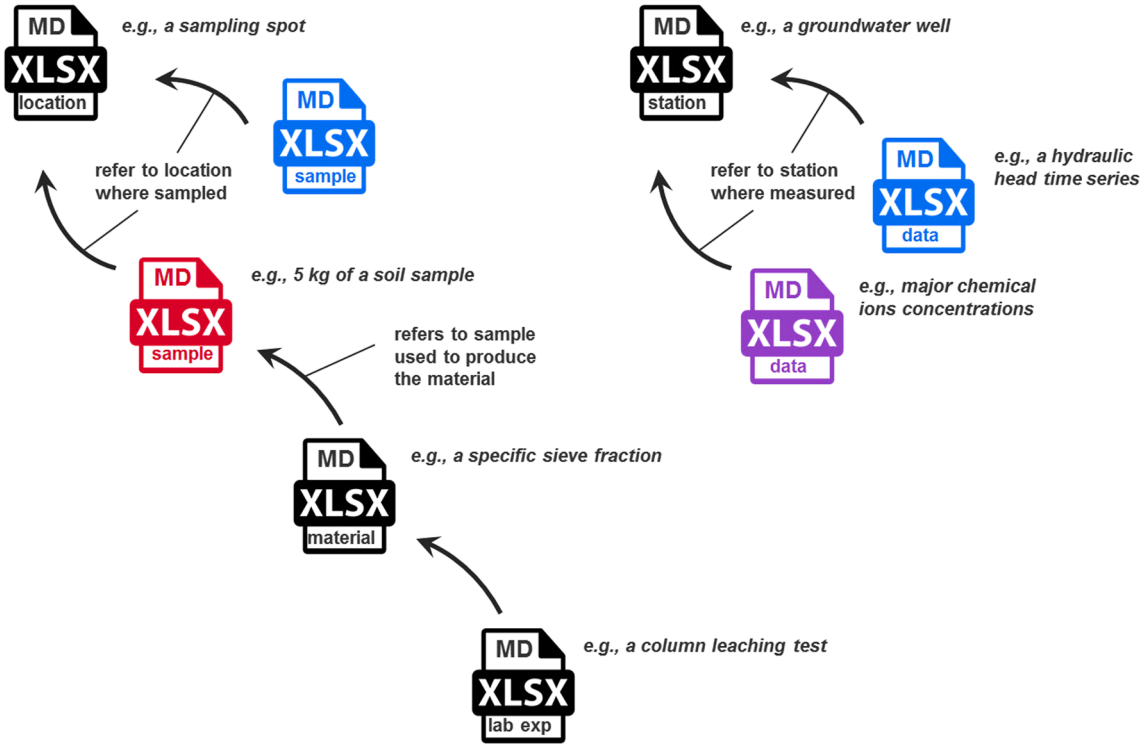

Fig. 3 Concept of hierarchical metadata (here taking the example of field measurement and sampling data) that reference (i.e. link) to respective metadata on higher hierarchy levels (icons modified based on icons made by Freepik from www.flaticon.com)

jet but is possible in Fedora-based repositories (Razum et al. 2007) and might be introduced in future FDAT releases.

Ingested metadata may be exported in other file formats such as Dublin Core XML records (https://dublincore.org) and in MARC 21 (https://www.loc.gov/marc/bibliographic). FDAT also provides the protocol for metadata harvesting (PMH) developed by the Open Archives Initiative (OAI) (http://www.openarchives.org/pmh/) to expose its contents to external services (Kaminski and Brandt 2018). FDAT is certificated by CoreTrustSeal (https://www.coretrustseal.org/) and registered at re3data.org (2018). The data structure in the long-term data archive FDAT is organized hierarchically in data containers. Each container has a certain number of members, which may be data sets or sub containers. Both containers and data sets get a unique persistent object identifier. A similar data repository design has been implemented at the Imperial College London (Harvey et al. 2017). This way we enable crediting of shared data for individual data sets as well as for collections of data stored in respective containers according to standard practice (Smith et al. 2016; Martone 2014).

Public web portal

The Public Web Portal is a web application that interfaces directly with CAMPOS-internal resources. It is designed as a starting point for an extendable platform providing further visualization and processing capabilities for the public or non-CAMPOS users. Please note that this part of the data management framework has not been implemented yet (see also section 4).

Researchers' workflow: From data and metadata creation to long-term preservation and retrieval

Overview

Both scientists and technicians involved in research take an active role in research data management. Their dedication to a prompt and proper storage of data and annotation by metadata is of utmost importance. Without their continuous contribution, research data will be insufficiently managed, either in terms of time (data resides too long in the researcher's workplace environment and is therefore not taken into account in the integrated analysis), documentation (metadata is incomplete to ensure that data can be searched and fully understood), or structure (structure of data does not meet basic requirements of any further processing or use of data). On these grounds we wish to illustrate the data management approach from the researchers' perspective describing their typical workflow from the generation and preparation of data and metadata to the long-term preservation of data including metadata. An overview of the most important steps of this workflow for data management is given in Fig. 4.

The main tasks, data preparation, metadata creation, maintenance, and long-term preservation of data are described in more detail below.

Data preparation

The introduction of collaborative data management does not necessarily mean additional work in data preparation. If researchers' preparation of data follows common rules of good practice in scientific data preparation (e.g., European Environment Agency 2012; Van den Eynden et al. 2011), the data can be stored in the Central Storage right away. The preparation of data refers not only to the data in a narrower sense, but also to any supplementary data or information that describes or explains the data.

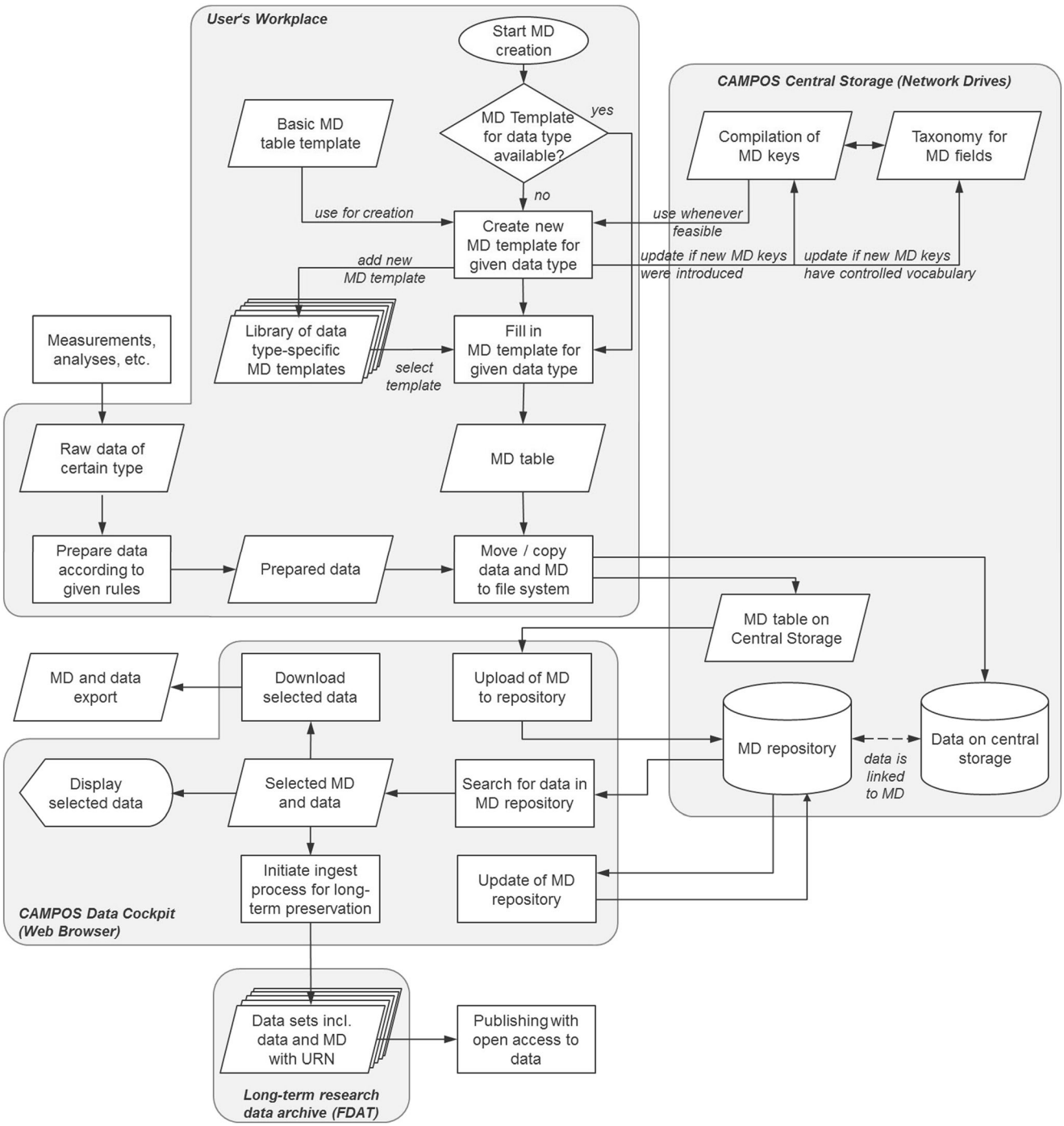

Fig. 4 Researchers' workflow scheme for data and metadata (MD) creation and storage

Metadata creation

To ease the creation of required metadata, metadata templates (spreadsheet files) are defined for each data type. Each of these templates is supposed to best describe the specific type of data in its respective context. The amount of effort required by the users depends on whether a metadata template for the given type of data is already available in the library of templates. This library can be accessed directly or via the web browser in the CAMPOS Data Cockpit. If a suitable template already exists the metadata creation just consists of making a copy of the template and filling in the metadata fields (see examples of metadata template tables in Figs. S1 to S5). Otherwise a new metadata template needs to be defined which might be useful for other users later. The number of available templates will grow during the course of a project, and the necessity to define new metadata templates will diminish over time. Each metadata set gets a unique identifier (ID), which is used to logically link metadata sets within the metadata structure (see Fig. 3). Descriptive IDs can be specified, which are checked for uniqueness during the automated upload validation (as described below).

All data type-specific metadata templates follow a predefined general template structure, which is organized in six different tables (Table 1). Each template is a compilation of information, which is required to adequately describe the respective data type. So-called metadata keys refer to individual items of information. New templates may be designed from scratch based on the general template structure, or, alternatively, an existing template may be modified. Only those sheets that include data type-specific metadata keys will be subject to modifications. Metadata keys should be chosen such that associated data is sufficiently described for any further use. The provided information must cover data generation, content, context, quality, structure, accessibility, and an update schedule of the data. Metadata keys should preferably be selected from the latest compilation of all metadata keys used within existing metadata templates. This is to keep the number of different metadata keys manageable. If none of the previously defined metadata keys is appropriate for the type of information that needs to be documented, a new metadata key has to and can be introduced, and the compilation of all metadata keys updated. For any newly introduced metadata key, the responsible researcher needs to decide whether a controlled vocabulary should be assigned or not (see, e.g., Hedden (2010) for some rationales). The template tables also include columns for the definition of the metadata type (string, text, integer, float, boolean), to indicate whether a metadata field input is mandatory or optional, and for the explanation of each metadata key to guide the researcher in using the template. The finalized new metadata template including possible extensions of the taxonomy is reviewed by the data team and taxonomists group and registered by the project data manager. After registration, the newly created template is added to the library of metadata templates. Corresponding metadata upload validation checks are automatically integrated in the Data Cockpit. If necessary, the taxonomy is updated. From there on the template is available for further use by all CAMPOS project members.

Data storage for internal access, maintenance, publication, and long-term preservation

After the preparation is completed, all metadata and data files are transferred from the user's workplace to the file system of the central data storage of the project. The file system has a default structure for higher folder levels (see Table S1) to make sure certain overarching principles are obeyed, such as the principle to structure data according to their locations (sub catchment, plot, station) and the environmental compartment to which they are related to (soil, surface water, groundwater, land surface, etc.). This is to reflect and ease the integrated approach of data analysis and interpretation on file system level, which would be significantly more difficult, if data were structured according to discipline or sub project. Only the project data managers (PDMs) have the permission to make modifications and additions at higher folder levels. The structure of lower folder levels is essentially free and every project member may modify and add folders in coordination with the responsible PDM. The data files that are transferred from the user's workspace to the central storage may also include supplementary files (see above).

To make data searchable and the access easier for all project members, metadata are then uploaded to the metadata repository of CAMPOS. This upload, as well as all further actions related to the maintenance, long-term preservation, and publication of metadata and data, is done with the help of functionalities provided by the CAMPOS Data Cockpit web interface (Fig. 5). As part of the metadata upload process, metadata are automatically validated (via checks answering whether all mandatory fields filled-in, and all used terms do follow the taxonomy of terms). Links between metadata and data files are generated based on the information given in the sheet FileDescription of the metadata (see Table 1).

After upload, the metadata is stored in the CAMPOS metadata repository. Any further modification of the metadata set is done using the Data Cockpit, in which metadata sets can be easily searched and maintained, and edited (see Fig. S6). Data search and access are permitted to all members. Permission to update is with the owner of the metadata set and possibly further project members – as granted and assigned by the owner. As mentioned above, the data itself remains on the file system and may be subject to regular updates.

Table 1 General structure of metadata (references are given to figures in the supplementary material)

Name of table sheet as defined in the general template	Content	Examples of metadata keys in particular sheet	User action in defining a new template
Main	Fixed set of main metadata keys (see Fig. S1).	NumberOfFiles MDsetCreatorSurName DatasetOwner	None
IndividualFields	Datatype-related metadata keys (see examples in Fig. S2).	RelatedCompartment SurfaceAltitude DrillingMethod SamplingMode	Define metadata keys as appropriate; take keys from latest compilation of metadata keys where possible
ColumnDescription	Metadata keys describing the content of individual data columns – if applicable to data (see example in Fig. S3).	Medium Parameter UnitOfMeasure MissingDataCode	Define metadata keys as appropriate; take keys from latest compilation of metadata keys where possible
FileDescription	Metadata keys to specify name, location, type and content of related (supplementing or documenting) files (see example in Fig. S4).	FileName Format CreatorSurName	None
ExtensionMetadatasets	Contains a list of one or more IDs to link the current metadata to other metadata sets - if applicable.	No metadata keys used in this table sheet	None
ControlledVocabulary	Lists of flat controlled vocabularies, i.e. pick lists, to control possible inputs for specific metadata keys (see example in Fig. S5).	No metadata keys used in this table sheet	Add pick lists for newly introduced metadata keys where appropriate

The way research data including metadata is managed in the CAMPOS-Internal area, makes it directly ready for publication. To do so, the researcher (typically the metadata owner) must first decide for an appropriate data repository. The Data Cockpit supports the researcher in transferring the data to the chosen repository by an automated bundling of proper data

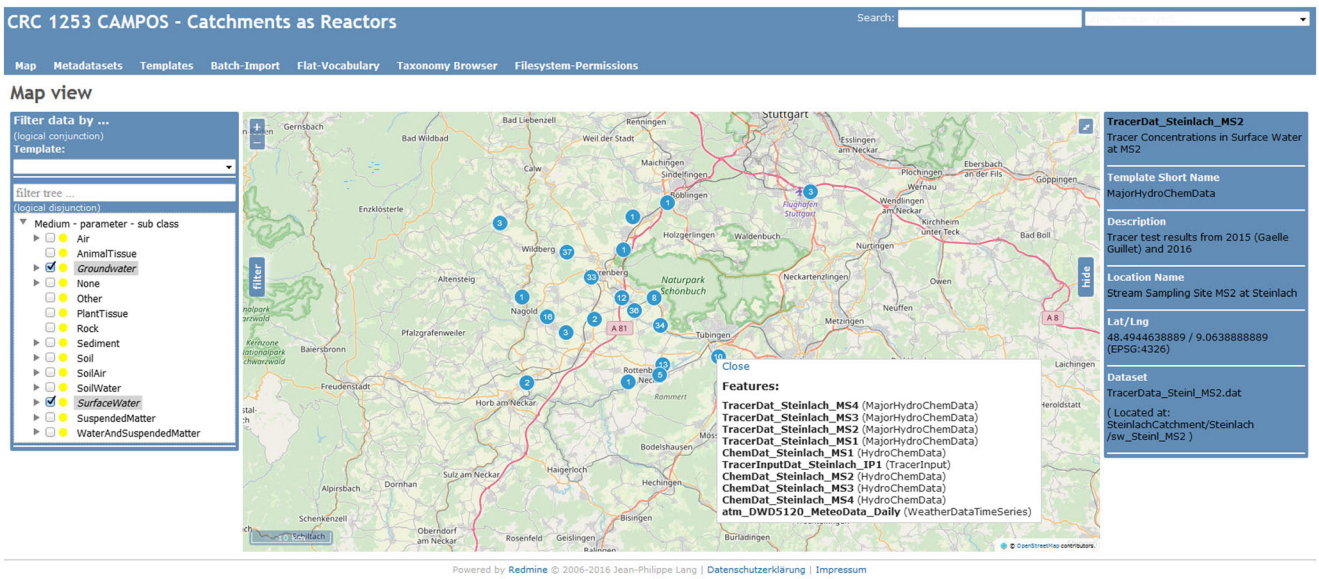

Fig. 5 CAMPOS Data Cockpit: map view on selected data

packages. Currently, support is available only for data transfer to the university's research data portal FDAT but this functionality is planned to be extended to support the data transfer also to other, discipline-specific repositories, which are recognized by the particular community. By initiating the archival and publication process, the data, metadata, and (optionally) supplementing files are automatically bundled to an ingest package and send to the archive. The package includes also information about the selected Creative Commons (CC) license (https://creativecommons.org) following a recommendation of the German Research Foundation (DFG 2014). The CAMPOS steering committee agreed to recommend a subset of CC licenses (CC-BY 3.0, CC-BY-SA 3.0 and CC-BY-ND 3.0), avoiding the non-commercial (NC) module, which might prevent access via free knowledge databases such as Wikipedia, open media archives and open source projects (Klimpel 2012).

Finally, the research data is checked by the FDAT staff members, assigned with a persistent identifier, and published in FDAT.

Implementing the data management framework: Time-line and experience

Almost three years after project launch, large parts of the research data management framework in CAMPOS are implemented. Although a few parts have not yet been completed (such as the implementation of the Live Web Portal) or need to be refined, we could put major parts of the framework into practice, as illustrated in the time-line of achievements (Fig. 6). The implemented data management infrastructure is increasingly being used by the researchers (> 80 registered users). This is the consequence of various factors, which have facilitated a rather smooth implementation and use of the research data management:

- The sufficient allocation of personnel resources (two positions, several student assistants) was a fundamental requirement for ensuring, one the one hand, a continuous development, technical implementation and administration of the data management framework and, on the other hand, an adequate coordination, communication, and day-to-day support of researchers.
- The organizational measures taken made sure that the data management was regularly on the agenda on different working levels. This has led to a continuous and intensive communication about the preparation, storage, and publication of data and metadata within the individual research projects and across different projects.
- The specific consideration of existing research disciplines' and researchers' needs, namely the adoption of existing workflows and data documentation schemes, was the key to involve researchers in the development and implementation process at the earliest possible stage. The joint development of data-specific templates for metadata creation, following a researcher-curator collaborative approach, has been very successful. This is consistent with the findings of Rodrigues et al. (2019) who showed that researchers working on familiar datasets can contribute efficiently to the definition of metadata models, in particular to overcome the common problem of data descriptors lacking specificity.

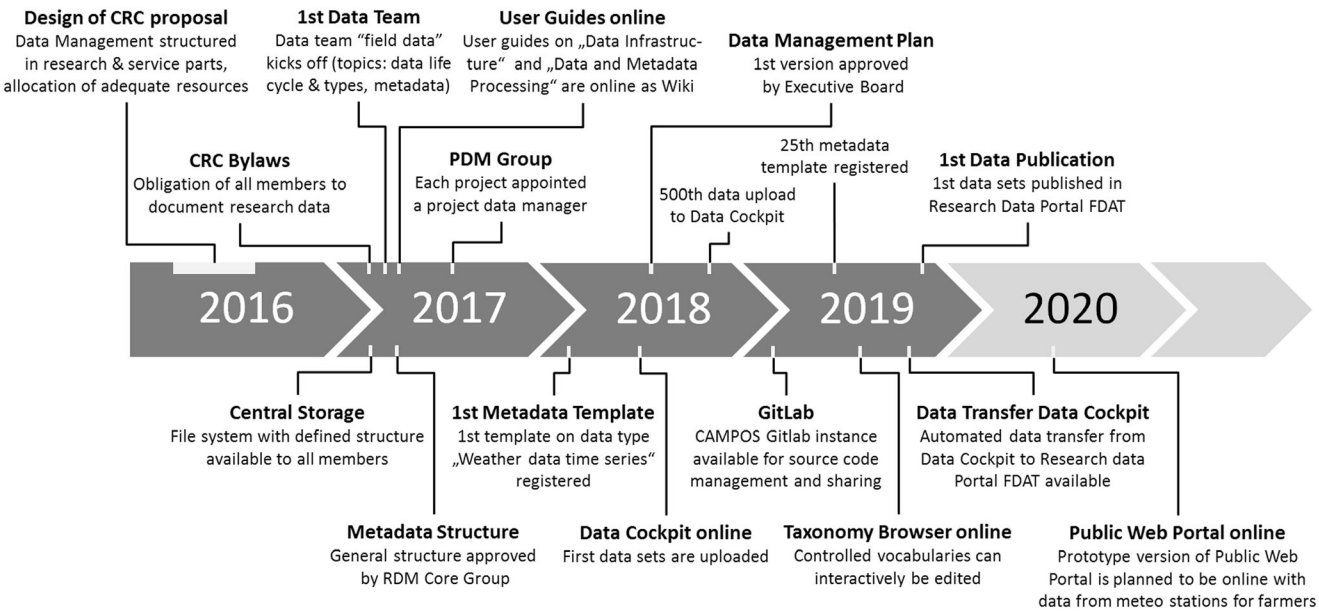

Fig. 6 Time line of the development and implementation of the CAMPOS data management framework

- The launch of the Central Storage, providing simple low-level data storage and exchange options, at an early stage of the project (Fig. 6) was an important step that helped to bring the data acquired in the various projects together on a common platform from the onset of the project, thus facilitating immediate collaborative use.
- Significant efforts to inform and train researchers in project-wide as well as focused seminars were essential to overcome the lack in understanding the requirements of research data management and were a prerequisite to allow for a distribution of activities and responsibilities among the CAMPOS researchers.
- Continuous endeavors to raise the researcher's awareness not only for the necessity of data management and sharing but also for its opportunities, together with the growing confidence in the suitability of the data management approach and its potential benefits (not least through the fast implementation of a first version of the Data Cockpit after approx. 18 months, see Fig. 6) were ultimately successful: in the third year of the project, the research data management in CAMPOS is understood as a joint task and responsibility as well as a collaborative opportunity for all project members.

Conclusions and outlook

We have described a collaborative research data management framework that was designed to meet the specific challenges of interdisciplinary and inter-institutional projects on integrated environmental research that are closely linked to the heterogeneity of the research data that vary in origin, type, scope, and size. The successful development and implementation of the framework within a relatively short period of three years, and the – to a great extent positive – experience made during this period with respect to the involvement of the researchers in the process of both development and implementation of the framework leads us to the conclusion that the underlying organizational and conceptual approach is feasible and recommendable, and may serve as a model for the setup of research data management platforms in other projects to come.

The implementation of the research data management framework will be continued and complemented by missing parts. Other parts will be refined or improved where necessary, in particular in those aspects where our experience has led us to identify room for improvements:

- Additional streamlining of the metadata creation process through electronic laboratory and field notebooks (e.g., Amorim et al. 2015) and tools for automated metadata creation (such as the generation of series of metadata of similar type) will further minimize the efforts required of the researchers.
- Enabling automated generation of user interfaces and database definitions (including validation) using a meta-description-framework for describing application internal model objects would reduce the maintenance effort significantly.
- Refactoring the application internal metadata representation with the help of the adaptive modeling pattern would increase the flexibility of metadata handling and mean an important step towards a more generic implementation that is capable of translating between a multitude of metadata standards.
- Extracting the implementation of the metadata application model, web services, and workflows into a more general (i.e. multipurpose) application plugin would be beneficial for other research communities as such a plugin could easily be added to already existing applications and will render own metadata tooling solutions unnecessary.

Further activities will be devoted to the development of data visualization and processing routines, the management of modelling data, and the embedding of the data in an overarching research infrastructure that allows coherent linking and combining data with data collected in other databases in a broader framework. To develop such a framework, a concerted action is currently outlined by a large number of research institutions in the field of earth system sciences in Germany (https://www.nfdi4earth.de/). Particular goals of NFDI4Earth, which is planned to be connected to the European Open Science Cloud (EC 2017), are (i) the development of a common conception and architecture, (ii) the technical development and establishment of usable, subject-specific research data workflows and trustworthy data services, (iii) the sustainable staffing of institutions with developers and data specialists, and (iv) the "digital qualification" of researchers.

Funding Information Open Access funding provided by Projekt DEAL. This work was supported by the Collaborative Research Center 1253 CAMPOS, funded by the German Research Foundation (DFG, Grant Agreement SFB 1253/1 2017).

References

Amorim R.C., Castro J.A., da Silva J.R., Ribeiro C. (2015) Engaging Researchers in Data Management with LabTablet, an Electronic Laboratory Notebook. In: Sierra-Rodríguez JL., Leal JP., Simões A. (eds) Languages, Applications and Technologies (SLATE) 2015. Communications in Computer and Information Science 563, Springer, https://doi.org/10.1007/978-3-319-27653-3_21

Amorim RC, Castro JA, Rocha da Silva J, Ribeiro C (2017) A comparison of research data management platforms: architecture, flexible metadata and interoperability. Univ Access Inf Soc 16:851–862. https://doi.org/10.1007/s10209-016-0475-y

Berlin declaration on open access to knowledge in the sciences and humanities (2003), https://openaccess.mpg.de/Berlin-Declaration, 2003 (Last Accessed Aug 2019)

Blaylock BK, Horel JD, Liston ST (2017) Cloud archiving and data mining of high-resolution rapid refresh forecast model output. Comput Geosci 109:43–50. https://doi.org/10.1016/j.cageo.2017.08.005

Chunpir, H.I. (2018) How to Include Users in the Design and Development of Cyberinfrastructures? In: Marcus, A., W. Wang, W. (eds.) Design, User Experience, and Usability (DUXU) 2018, Lecture notes in computer science (LNCS) 10918, Springer, 658–672, https://doi.org/10.1007/978-3-319-91797-9_46

Curdt, C. (2016) Metadata Management in an Interdisciplinary, Project-Specific Data Repository: A Case Study from Earth Sciences. In: Garoufallou E., Subirats Coll I., Stellato A., Greenberg J. (eds) Metadata and Semantics Research. MTSR 2016. Communications in Computer and Information Science 672, Springer, https://doi.org/10.1007/978-3-319-49157-8_31

Curdt C (2019) Supporting the interdisciplinary, long-term research project 'patterns in soil-vegetation-atmosphere-systems' by data management services. Data Science Journal 18(1):1–9. https://doi.org/10.5334/dsj-2019-005

Dehnhard I, Weichselgartner E, Krampen G (2013) Researcher's willingness to submit data for data sharing: A case study on a data archive for psychology. Data Science Journal, Data Science Journal 12:172–180. https://doi.org/10.2481/dsj.12-037

DFG (2014) Deutsche Forschungsgemeinschaft (German Research Foundation): Information für die Wissenschaft Nr. 68 (Information for Researchers No 68) Appell zur Nutzung offener Lizenzen in der Wissenschaft (Appeal for the use of open licenses in science) 20. November 2014, https://www.dfg.de/foerderung/info_wissenschaft/2014/info_wissenschaft_14_68/index.html. Accessed 2 Sept 2019

EC (2017) H2020 Programme Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020. Version 3.2 from 21 March 2017, European Commisssion, Directorate-General for Research & Innovation, 11 pp, http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf (accessed 08 July 2019)

European Environment Agency (2012) Prepare data for analysis and visualisations. European Environment Agency, Copenhagen, Denmark, 11 pp, https://www.eea.europa.eu/data-and-maps/daviz/learn-more/prepare-data (accessed 09 July 2019)

Fecher B, Friesike S, Hebing M (2015) What drives academic data sharing? PLoS One 10(2):e0118053. https://doi.org/10.1371/journal.pone.0118053

Geosling E, Pollak J, Hooper R (2015) Advancing water science through community collaboration. Environ Earth Sci 73:1919–1924. https://doi.org/10.1007/s12665-014-3835-z

Glatard T, Rousseau M-É, Camarasu-Pop S, Adalat R, Beck N, Das S, da Silva RF, Khalili-Mahani N, Korkhov V, Quirion P-O, Rioux P, Olabarriaga SD, Bellec P, Evans AC (2017) Software architectures to integrate workflow engines in science gateways. Future Generation Computer Systems, Future Generation Computer Systems 75:239–255. https://doi.org/10.1016/j.future.2017.01.005

Grunzke R, Hartmann V, Jejkal T, Kollai H, Prabhune A, Herold H, Deicke A, Dressler C, Dolhoff J, Stanek J, Hoffmann A, Müller-Pfefferkorn R, Schrade T, Meinel G, Herres-Pawlis S, Nagel WE (2019) The MASi repository service - comprehensive, metadata-driven and multi-community research data management. Futur Gener Comput Syst 94:879–894. https://doi.org/10.1016/j.future.2017.12.023

Harvey MJ, McLean A, Rzepa HS (2017) A metadata-driven approach to data repository design. Journal of Cheminformatics 9:4. https://doi.org/10.1186/s13321-017-0190-6

Hedden HJ (2010) Taxonomies and controlled vocabularies best practices for metadata. Digit Asset Manag 6:279–284. https://doi.org/10.1057/dam.2010.29

Horsburgh JS, Tarboton DG, Maidment DR, Zaslavsky I (2008) A relational model for environmental and water resources data. Water Resour Res 44. https://doi.org/10.1029/2007wr006392

Horsburgh JS, Tarboton DG, Hooper RP, Zaslavsky I (2014) Managing a community shared vocabulary for hydrologic observations. Environ Model Softw 52:62–73. https://doi.org/10.1016/j.envsoft.2013.10.012

Horsburgh JS, Reeder SL, Jones AS, Meline J (2015) Open source software for visualization and quality control of continuous hydrologic and water quality sensor data. Environ Model Softw 70:32–44. https://doi.org/10.1016/j.envsoft.2015.04.002

Hsu L, Martin RL, McElroy B, Litwin-Miller K, Kim W (2015) Data management, sharing, and reuse in experimental geomorphology: challenges, strategies, and scientific opportunities. Geomorphology, Geomorphology 244:180–189. https://doi.org/10.1016/j.geomorph.2015.03.039

Kaminski, S., Brandt, O. (2018) Das institutionelle Forschungsdatenrepositorium FDAT der Universität Tübingen. O-Bib. Das Offene Bibliotheksjournal 5(3), 61-75, VDB https://doi.org/10.5282/o-bib/2018H3S61-75

Klimpel, P. (2012) Freies Wissen dank creative-commons-Lizenzen: Folgen, Risiken und Nebenwirkungen der Bedingung „nicht-kommerziell"– NC (Free knowledge with creative commons licenses: consequences, risks and side effects of the condition "non-commercial" – NC). iRights.info, Berlin, https://irights.info/wp-content/uploads/userfiles/CC-NC_Leitfaden_web.pdf ()

Kratz JE, Strasser C (2015) Researcher perspectives on publication and peer review of data. PLoS One 10(2):e0117619. https://doi.org/10.1371/journal.pone.0117619

Latham B (2017) Research data management: defining roles, prioritizing services, and enumerating challenges. J Acad Librariansh 43:263–265. https://doi.org/10.1016/j.acalib.2017.04.004

Martone, M. (2014) Data citation synthesis group: joint declaration of data citation principles, FORCE11, https://doi.org/10.25490/a97f-egyk

Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, Buck S, Chambers CD, Chin G, Christensen G, Contestabile M, Dafoe A, Eich E, Freese J, Glennerster R, Goroff D, Green DP, Hesse B, Humphreys M, Ishiyama J, Karlan D, Kraut A, Lupia A, Mabry P, Madon T, Malhotra N, Mayo-Wilson E, McNutt M, Miguel E, Paluck EL, Simonsohn U, Soderberg C, Spellman BA, Turitto J, VandenBos G, Vazire S, Wagenmakers EJ, Wilson R, Yarkoni T (2015) Promoting an open research culture. Science 348:1422–1425. https://doi.org/10.1126/science.aab2374

Piasecki M, Beran B (2009) A semantic annotation tool for hydrologic sciences. Earth Sci Inform 2:157–168. https://doi.org/10.1007/s12145-009-0031-x

Pinfield, S., Cox, A., Smith, J.R. (2014) Research data management and Libraries: Relationships, Activities, Drivers and Influences. PLOS ONE, https://doi.org/10.1371/journal.pone.0114734

Razum, M., Schwichtenberg, F., Fridman, R. (2007) Versioning of Digital Objects in a Fedora-based Repository. http://hdl.handle.net/11858/00-001M-0000-0013-B0CF-5

re3data.org (2018) Research Data Portal FDAT. Editing status 2018-10-02, re3data.org - Registry of Research Data Repositories, https://doi.org/10.17616/R3PM1K

Rodrigues J., Castro J.A., da Silva J.R., Ribeiro C. (2019) Hands-On Data Publishing with Researchers: Five Experiments with Metadata in Multiple Domains. In: Manghi P., Candela L., Silvello G. (eds) Digital Libraries: Supporting Open Science. IRCDL 2019. Communications in Computer and Information Science 988, Springer, https://doi.org/10.1007/978-3-030-11226-4_22

Sadler JM, Ames DP, Khattar R (2016) A recipe for standards-based data sharing using open source software and low-cost electronics. J Hydroinf 18:185–197. https://doi.org/10.2166/hydro.2015.092

Smith, A.M., Katz, D.S., Niemeyer, K.E. (2016) FORCE11 Software Citation Working Group. 2016. Software citation principles. PeerJ Computer Science 2:e86, https://doi.org/10.7717/peerj-cs.86

Specht A, Guru S, Houghton L, Keniger L, Driver P, Ritchie EG, Lai K, Treloar A (2015) Data management challenges in analysis and synthesis in the ecosystem sciences. Sci Total Environ 534:144–158. https://doi.org/10.1016/j.scitotenv.2015.03.092

Tenopir C, Dalton ED, Allard S, Frame M, Pjesivac I, Birch B, Pollock D, Dorsett K (2015) Changes in data sharing and data reuse practices and perceptions among scientists worldwide. PLoS One 10(8): e0134826. https://doi.org/10.1371/journal.pone.0134826

Van den Eynden, V., Corti, L., Woollard, M., Bishop, L., Horton, L. (2011) Managing and sharing data – best practice for researchers. UK data archive, 3rd edition, University of Essex, may 2011, 40 pp.

Wang, W.M., Göpfert, T., Stark, R. (2016) Data management in collaborative interdisciplinary research projects - conclusions from the digitalization of research in sustainable manufacturing. ISPRS International Journal of Geo-Information, ISPRS International Journal of Geo-Information 5, https://doi.org/10.3390/ijgi5040041

White HC (2014) Descriptive metadata for scientific data repositories: A comparison of information scientist and scientist organizing behaviors. J Libr Metadata 14(1):24–51. https://doi.org/10.1080/19386389.2014.891896

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B. (2016) The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3, doi:https://doi.org/10.1038/sdata.2016.18

Wilkinson MD, Sansone S-A, Schultes E, Doorn P, da Silva Santos LB, Dumontier M (2018) A design framework and exemplar metrics for FAIRness. Scientific Data 5:1–4. https://doi.org/10.1038/sdata.2018.118