EGU2020: Sharing Geoscience Online Session
ESSI2.10 - Data Integration: Enabling the Acceleration of Science Through Connectivity, Collaboration, and Convergent Science
DOI of this presentation's display: 10.5194/egusphere-egu2020-8638

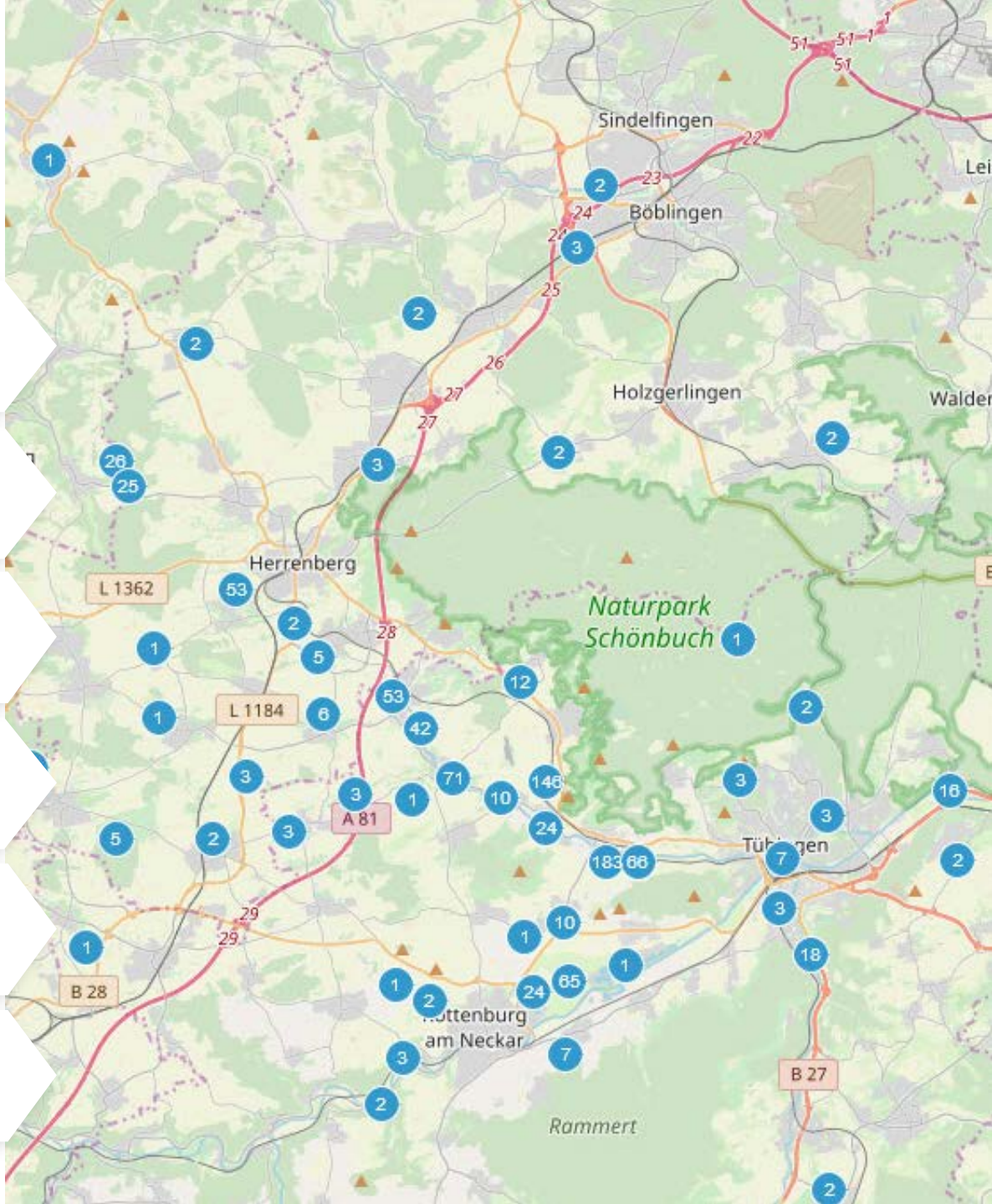
Managing collaborative research data for integrated, interdisciplinary environmental research

Michael Finkel, Albrecht Baur, Tobias K.D. Weber, Karsten Osenbrück, Hermann Rügner, Carsten Leven, Marc Schwientek, Johanna Schlögl, Ulrich Hahn, Thilo Streck, Olaf A. Cirpka, Thomas Walter, and Peter Grathwohl

See for more details:
Finkel, M., Baur, A., Weber, T. et al. Managing collaborative research data for integrated, interdisciplinary environmental research. Earth Sci Inform (2020). https://doi.org/10.1007/s12145-020-00441-0

Photography: Thomas Hentrich, www.MomentsInRGB.com

Overview

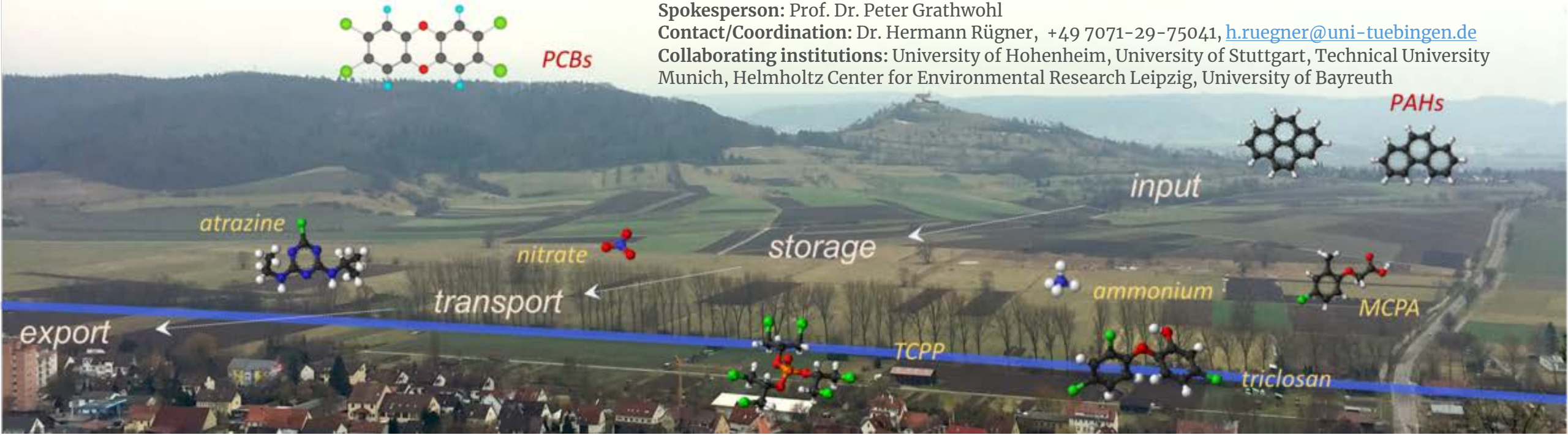


CRC 1253 CAMPOS - Catchments as Reactors
Metabolism of Pollutants on the Landscape Scale

The research in CAMPOS focuses on diffuse pollution of soils, surface waters, and groundwater by a multitude of anthropogenic contaminants and their turnover at landscape scale. CAMPOS 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 being produced.

The spectrum of research data ranges from concentration measurements in water and soil, high-resolution molecular fragments data, toxicity test data, hydrological, geological, and hydrogeological data, to genomic and metabolomics data from molecular-biological analysis. Produced data sets differ in terms of size, dimension, structure, format, temporal frequency, and origin, amongst others. The data are used to inform/calibrate numerical models to simulate water fluxes and reactive solute transport.

Spokesperson: Prof. Dr. Peter Grathwohl
Contact/Coordination: Dr. Hermann Rügner, +49 7071-29-75041, h.ruegner@uni-tuebingen.de
Collaborating institutions: University of Hohenheim, University of Stuttgart, Technical University Munich, Helmholtz Center for Environmental Research Leipzig, University of Bayreuth



Organizational measures for data management

Focus on data management from the beginning

Building awareness of data management requirements and goals within the consortium from the beginning of the project proposal activities

In the proposal:

- emphasise the importance of data management for the project success
- commitment to concertedly conceptualize & develop procedures / tools to implement the integrated data management

Data management as research and service

Work is structured into

- a research part (development of concepts and infrastructure), and
- a service part (analysis of researchers' demands, coordination of implementation and operation, training)

Maximum communication

Premise for involvement of the entire consortium in data management activities

Meetings of Core Group, data teams, seminars, workshops, etc.

Appropriate organizational entities

Organizational entities on three levels to efficiently coordinate the development and implementation of the data management within the CRC:

- 'Executive Board': top level decisions – strategy, prioritization, licensing, etc.
- 'Project Data Managers' (PDMs): organization of work in individual projects;
- 'Data Teams': management of specific types of data, incl. metadata definitions

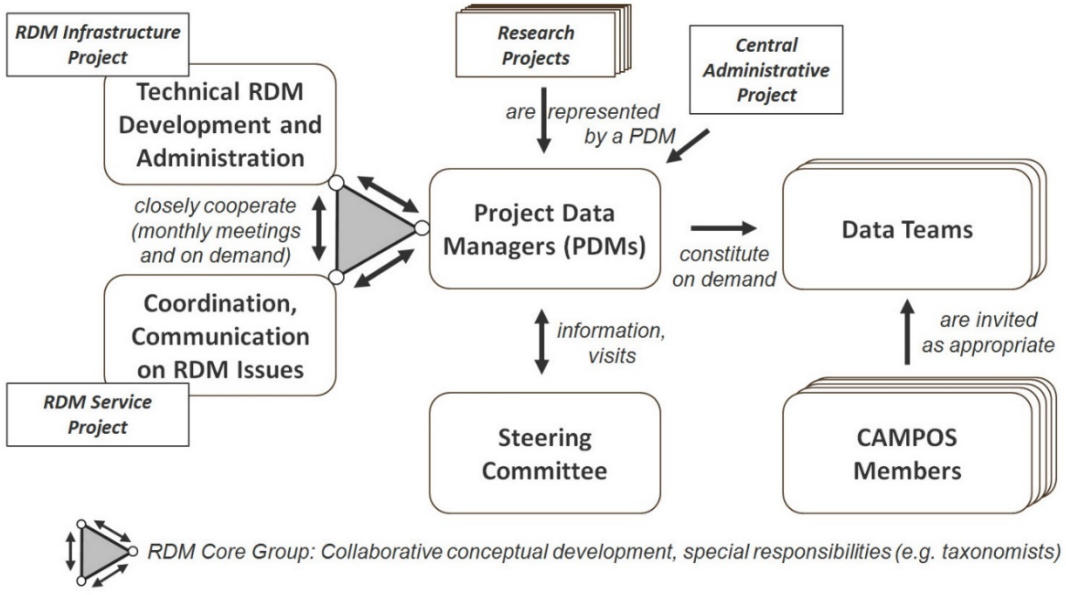


Fig. 1 Organizational entities for the conceptualization, development and implementation of data management within CAMPOS (Source: Finkel et al. 2020)

Core elements and structure of data management

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 (see Figure on the right) based on two core conceptual elements.

Separation of ongoing research data vs. publicly available data

- the CAMPOS-internal area provides a 'safe environment' for all CAMPOS research data
- the Public Web Portal and the Research Data Archive allow public access

Separation of data and metadata

- Data is stored in a file system (CAMPOS Central Storage)
- Metadata is managed in a database (CAMPOS Metadata Repository)

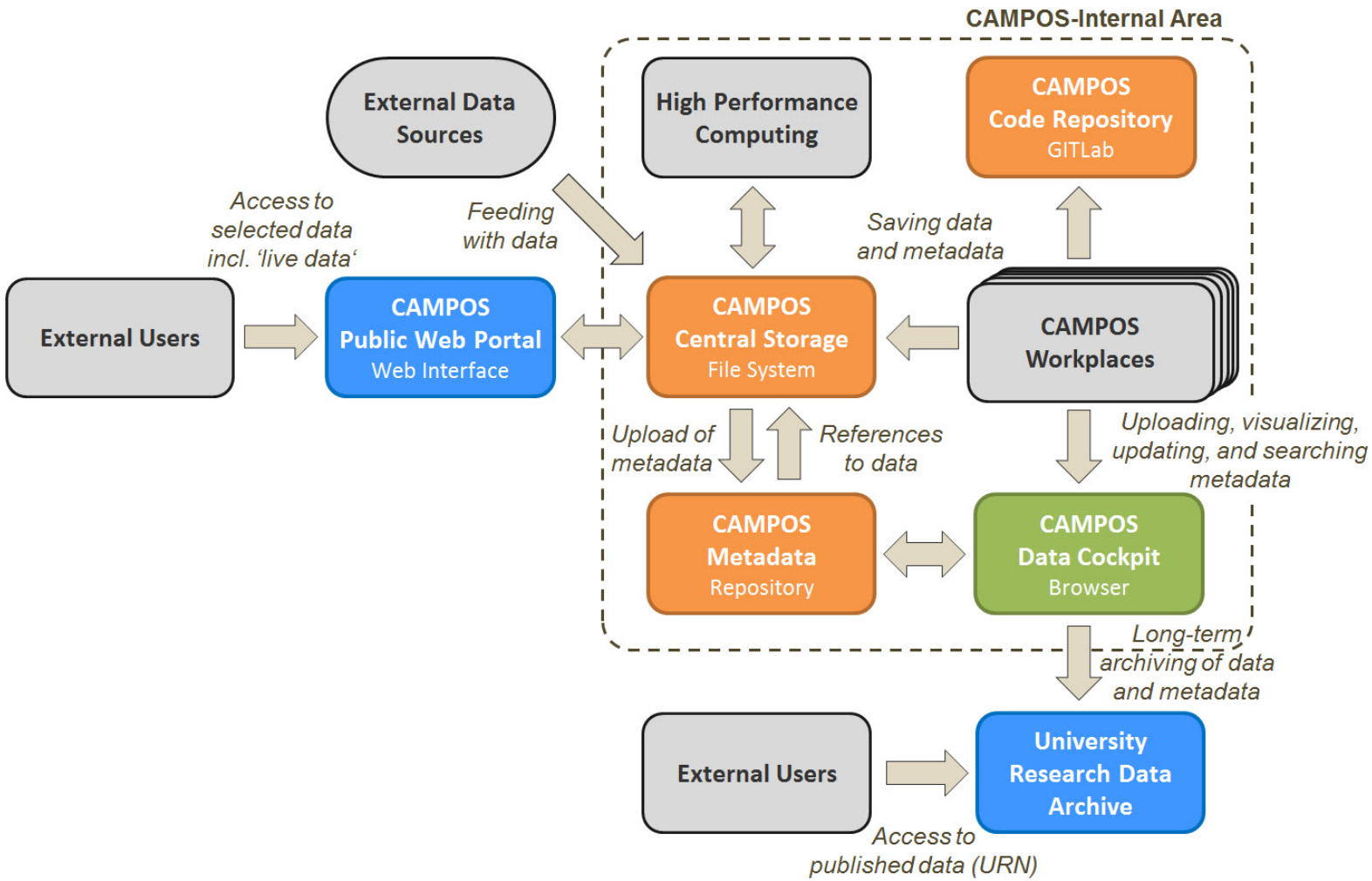


Fig. 2 Structure of the CAMPOS data management framework (Source: Finkel et al. 2020)

Hierarchical structure of metadata

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.

This concept offers large flexibility and efficiency because metadata can be defined in a data type-specific way.

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 datamanagement can be restricted to their particular research data and description of procedures, minimizing their efforts by avoiding any additional, unnecessary expense.

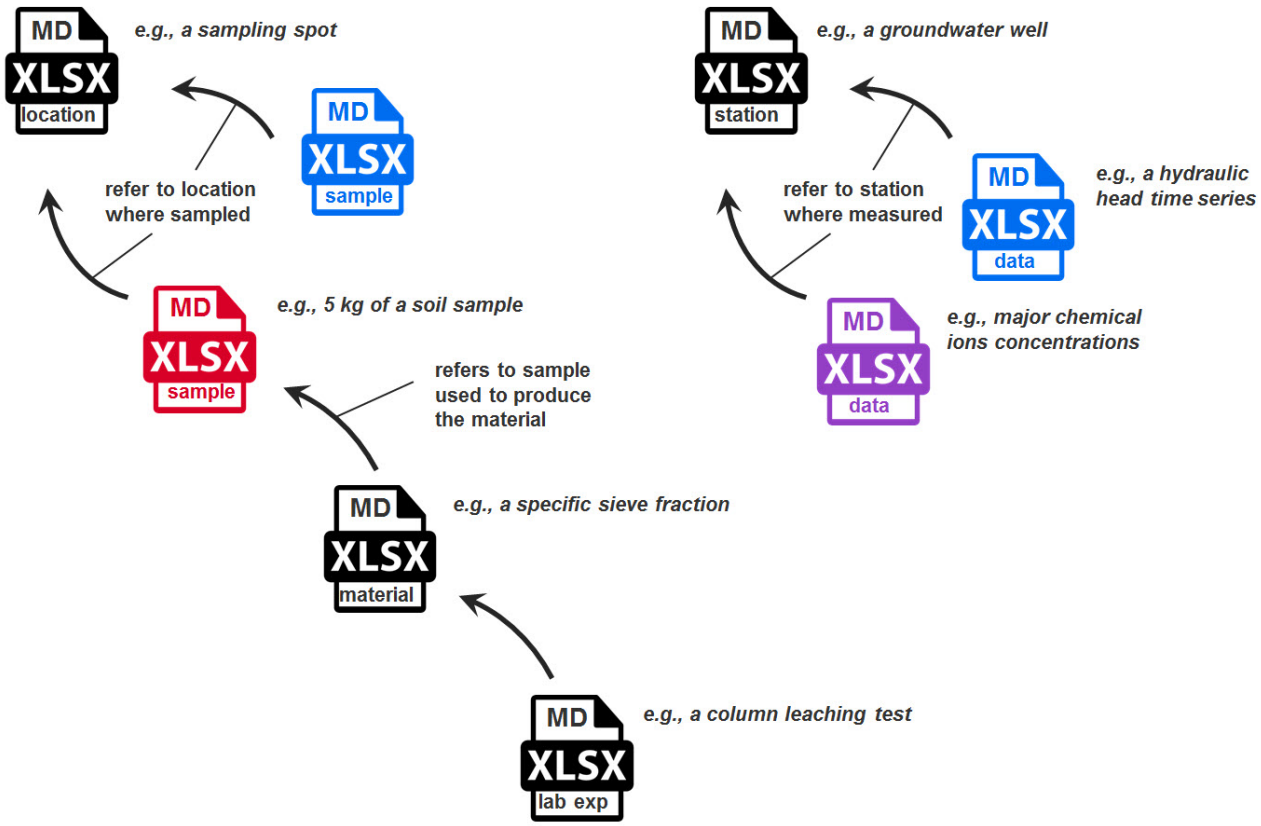


Fig. 3 Concept of hierarchical metadata (here taking the example of field measurement and sampling data , respectively) that reference (i.e. link) to respective metadata on higher hierarchy levels (groundwater well and sampling spot, respectively). [Icons are modified based on icons made by Freepik from www.flaticon.com.] (Source: Finkel et al. 2020)

Data type-specific metadata templates

Metadata creation with templates following a common general metadata structure

Metadata is created by help of ***metadata templates*** (spreadsheet files) defined for each data type.

Each of these templates is supposed to best describe the specific type of data in its context.

All metadata templates follow a predefined general template structure (organized in six tables, Tab. 1) and are a compilation of information required to adequately describe the respective data type. So-called ***metadata keys*** refer to individual items of information.

The 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 (access via CAMPOS Data Cockpit) or not:

- if a suitable template already exists, metadata creation just consists of making a copy of the template and filling in the metadata fields
- otherwise a new template needs to be designed from scratch or by modification of an existing template.

Tab. 1 General structure of metadata (Source: Finkel et al. 2020)

Name of table sheet *)	Content	Examples of metadata keys
Main	Fixed set of main metadata keys	NumberOfFiles MDsetCreatorSurName DatasetOwner
IndividualFields	Datatype-related metadata keys	RelatedCompartment SurfaceAltitude DrillingMethod SamplingMode
ColumnDescription	Metadata keys describing the content of individual data columns	Medium Parameter UnitOfMeasure MissingDataCode
FileDescription	Metadata keys to specify name, location, type and content of related (supplementing or documenting) files	FileName Format CreatorSurName
ExtensionMetadatasets	Contains a list of one or more IDs to link the current metadata to other metadata sets	- None -
ControlledVocabulary	Lists of flat controlled vocabularies, i.e. pick lists, to control possible inputs for specific metadata keys	- None -

*) as defined in the general metadata structure

Metadata creation

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 using a standard spreadsheet-calculation software (as described above above). These metadata are transferred from the workplace of the respective researcher or technician to the file system of the CAMPOS Central Storage (formatted as OASIS OpenDocument or Office Open XML) – together with the corresponding data (see Fig. 4 – top).
- In a second step the metadata is uploaded into the database. This upload process includes an automated validation of the metadata (see Fig. 4 – bottom).

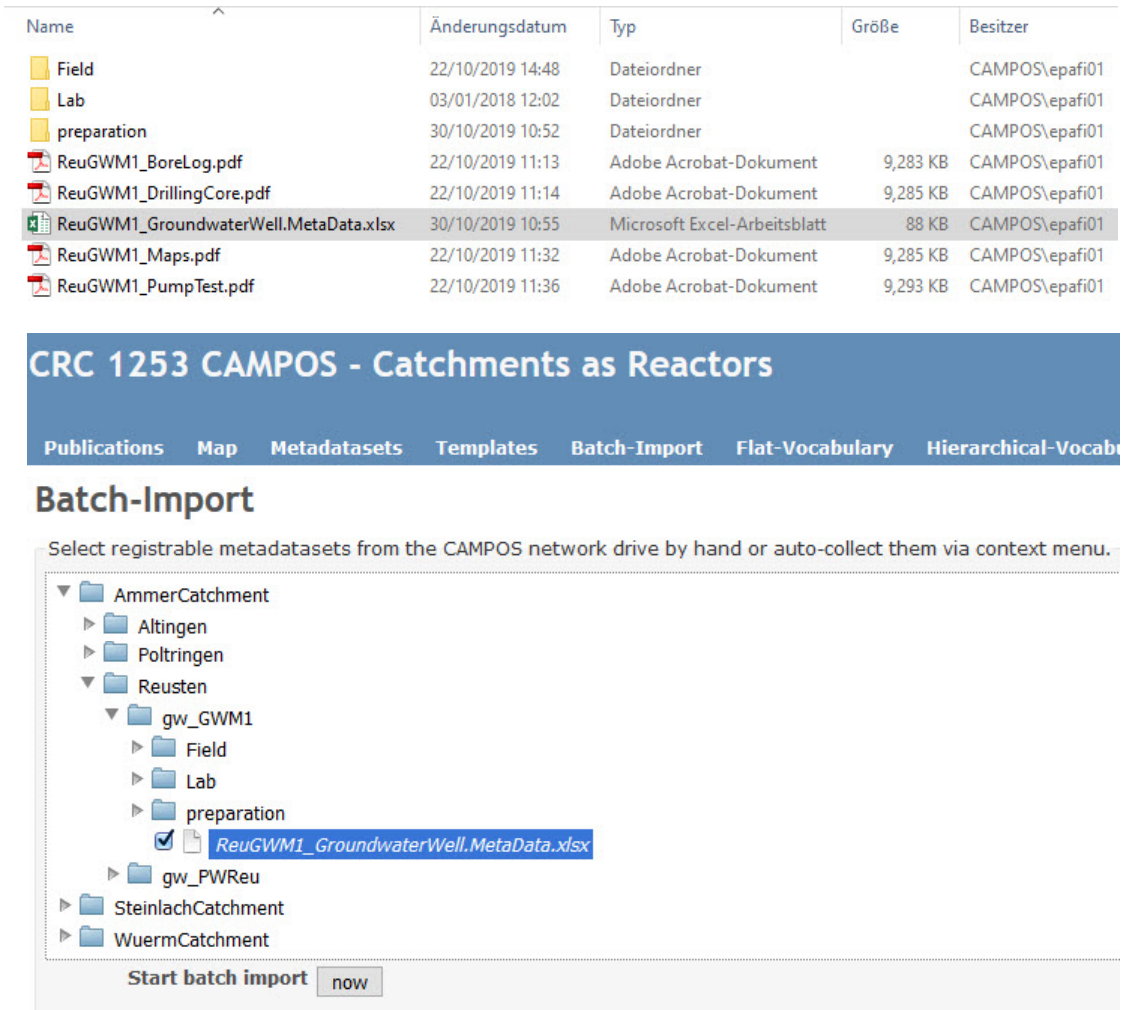


Fig. 4 Two-step metadata creation: preparation and transfer of metadata to file system (top), upload of metadata from/into CAMPOS Data Cockpit

Metadata creation

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.

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) to enable a flexible search and to overcome ontological and semantic heterogeneity when data is synthesized with other repositories.

All terms are in English. Controlled vocabularies are defined during the creation of data-type specific metadata templates if appropriate. Existing vocabularies may be adopted, e.g., if data from external sources is imported.

A central taxonomy service, as part of the CAMPOS Data Cockpit, offers convenient access to the vocabularies for all researchers and interactive editing capabilities for the taxonomists group to continuously update the vocabularies upon users' demand.

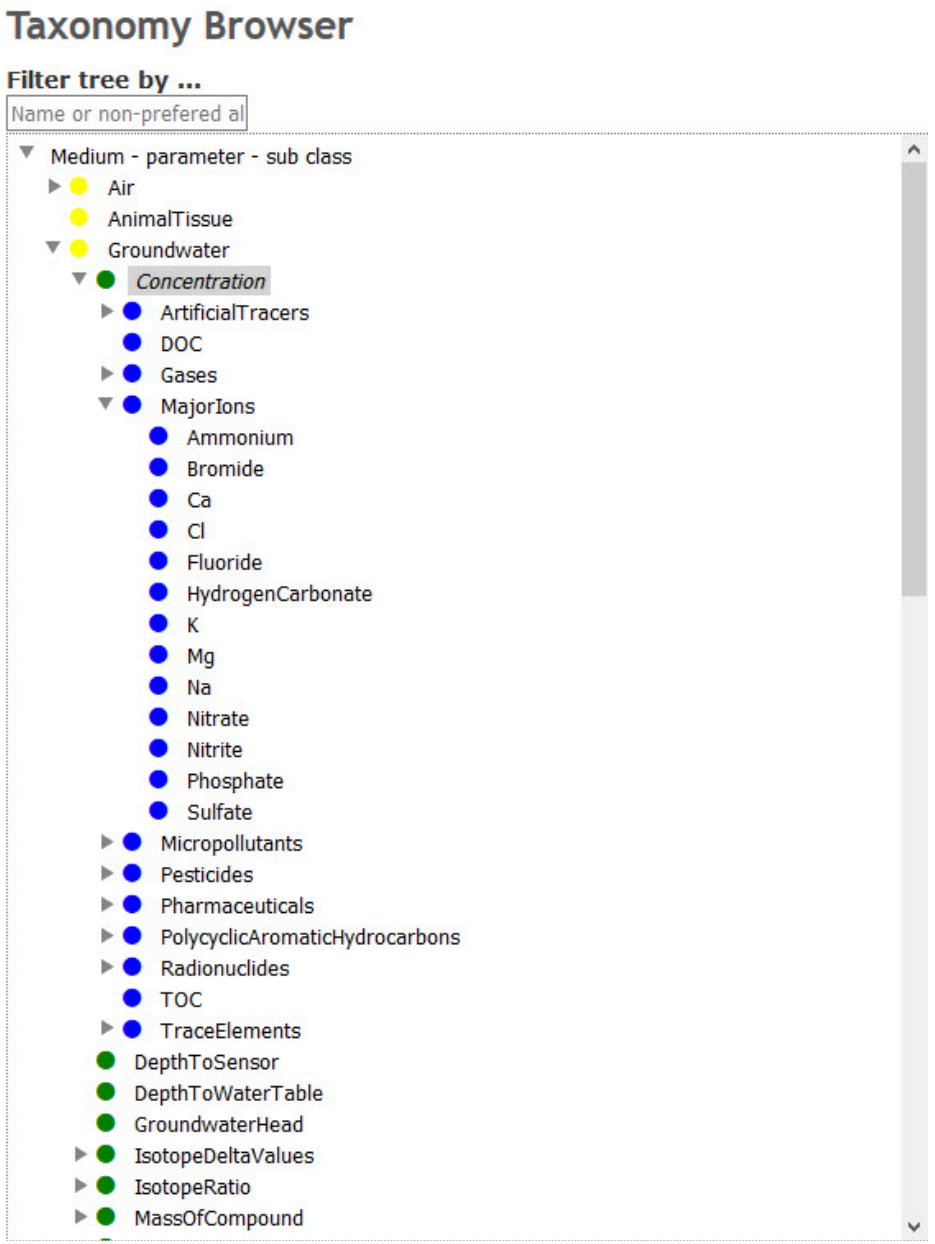


Fig. 5 Taxonomy browser of the CAMPOS Data Cockpit (detail)

Technical infrastructure – Functional environments

The CAMPOS framework for management of research data consists of three functional environments:

- the **CAMPOS Internal Area** forming the private working environment of the CAMPOS researchers for all data-related tasks and issues,
- the **Research Data Archive FDAT** of the University of Tübingen to preserve and publish data for long-term storage and use, and
- the **CAMPOS Public Web Portal** to provide public access to selected data (not yet implemented).

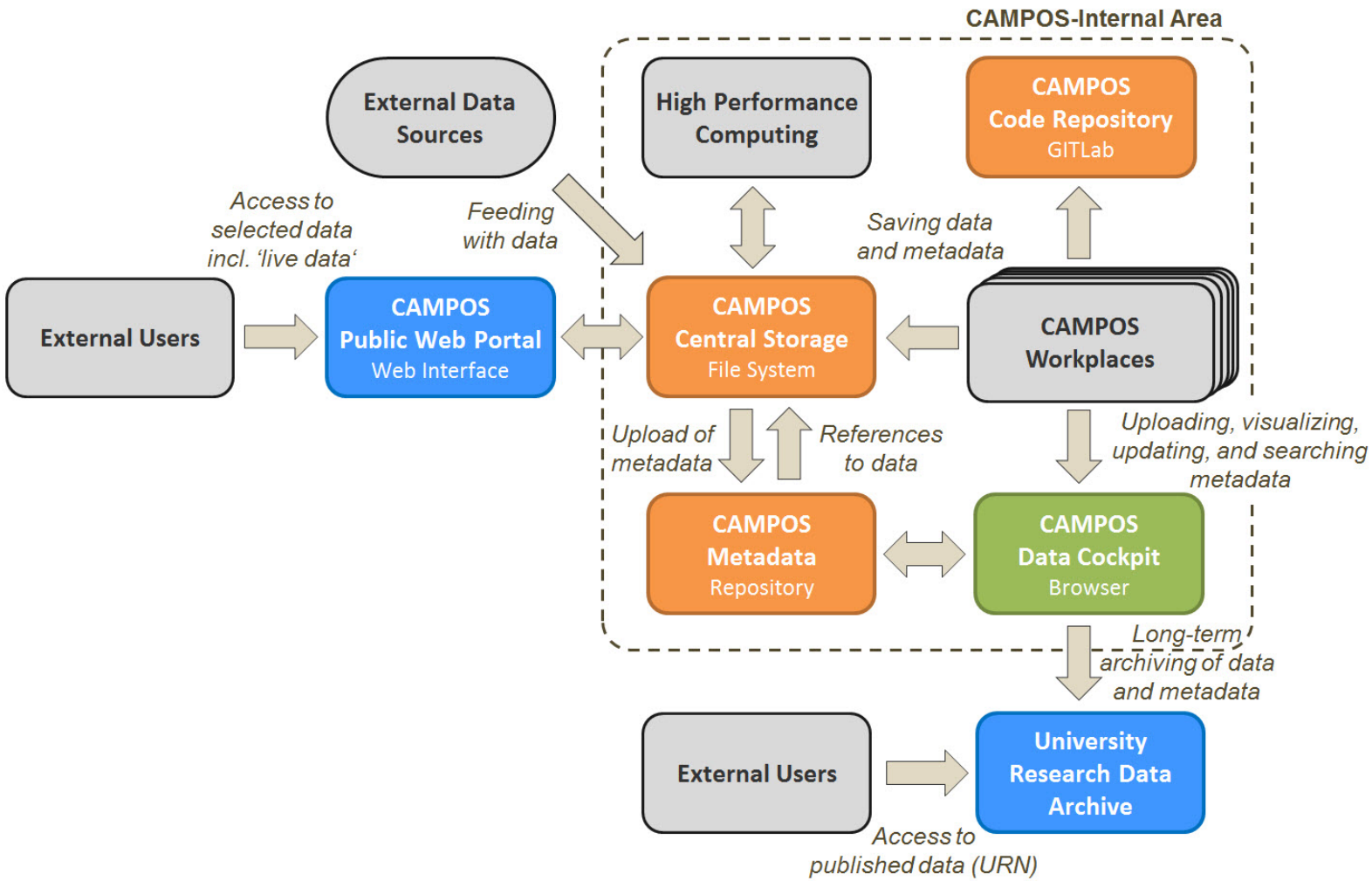


Fig. 2 Structure of the CAMPOS data management framework (Source: Finkel et al. 2020)

Technical infrastructure – CAMPOS internal area

The CAMPOS Internal Area consists of three main components:

CAMPOS Central Storage
File System

File system distributed as a SMB shared network drive (effective capacity of 100 TB) that stores all relevant internal research data of the CAMPOS project.

It is hosted at the central computing facilities of the University of Tübingen.

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.

CAMPOS Metadata
Repository

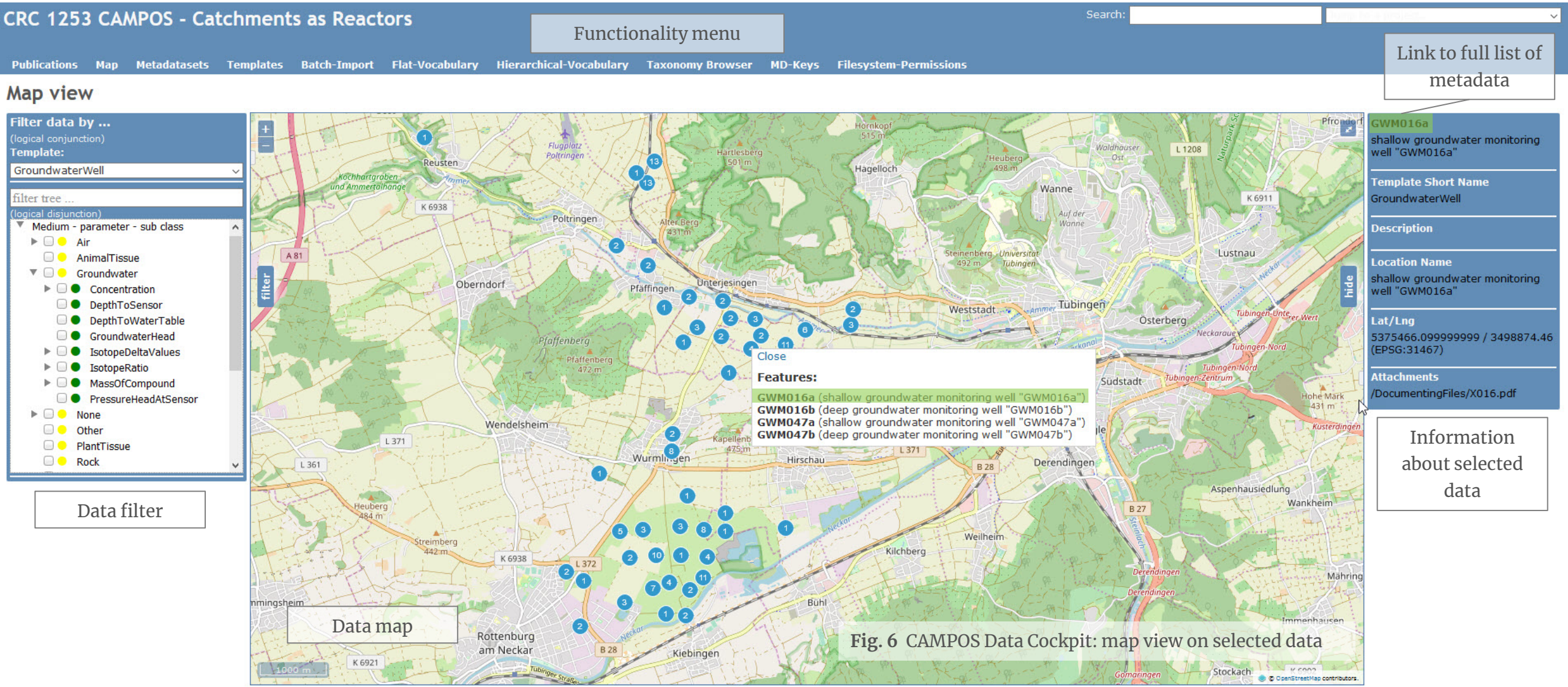
Relational database that holds all registered metadata sets including references to the actual data stored in the file system. It serves as the basis for data search.

CAMPOS Data Cockpit
Browser

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. It interfaces the Public Web Portal and the Research Data Archive FDAT.

See the next page for an illustration of the Data Cockpit.

Technical infrastructure – CAMPOS Data Cockpit



Powered by Redmine © 2006-2016 Jean-Philippe Lang

Technical infrastructure – Public areas

The CAMPOS data management also includes two public areas :

University Research Data Archive

The publication and archiving of data is accomplished via the university's long-term research data archive, the Research Data Portal FDAT (https://fdat.escience.uni-tuebingen.de/portal/).
The archive consists of the open source repository software Fedora Commons (https://duraspace.org/fedora), which is 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 data resources to be archived and FDAT-specific metadata required for ingest and archiving.

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).

A persistent, location-independent resource identifier, a uniform resource name (URN) is assigned to each data resource.

CAMPOS Public Web Portal
Web Interface

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 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.

Researchers' workflow

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

Both scientists and technicians involved in research take an active role in research data management. 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 it is important to consider 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 reservation of data including metadata. An overview of the most important steps of this workflow for data management is given in Fig. 7.

The main tasks are: data preparation, metadata creation, maintenance, and long-term preservation of data.

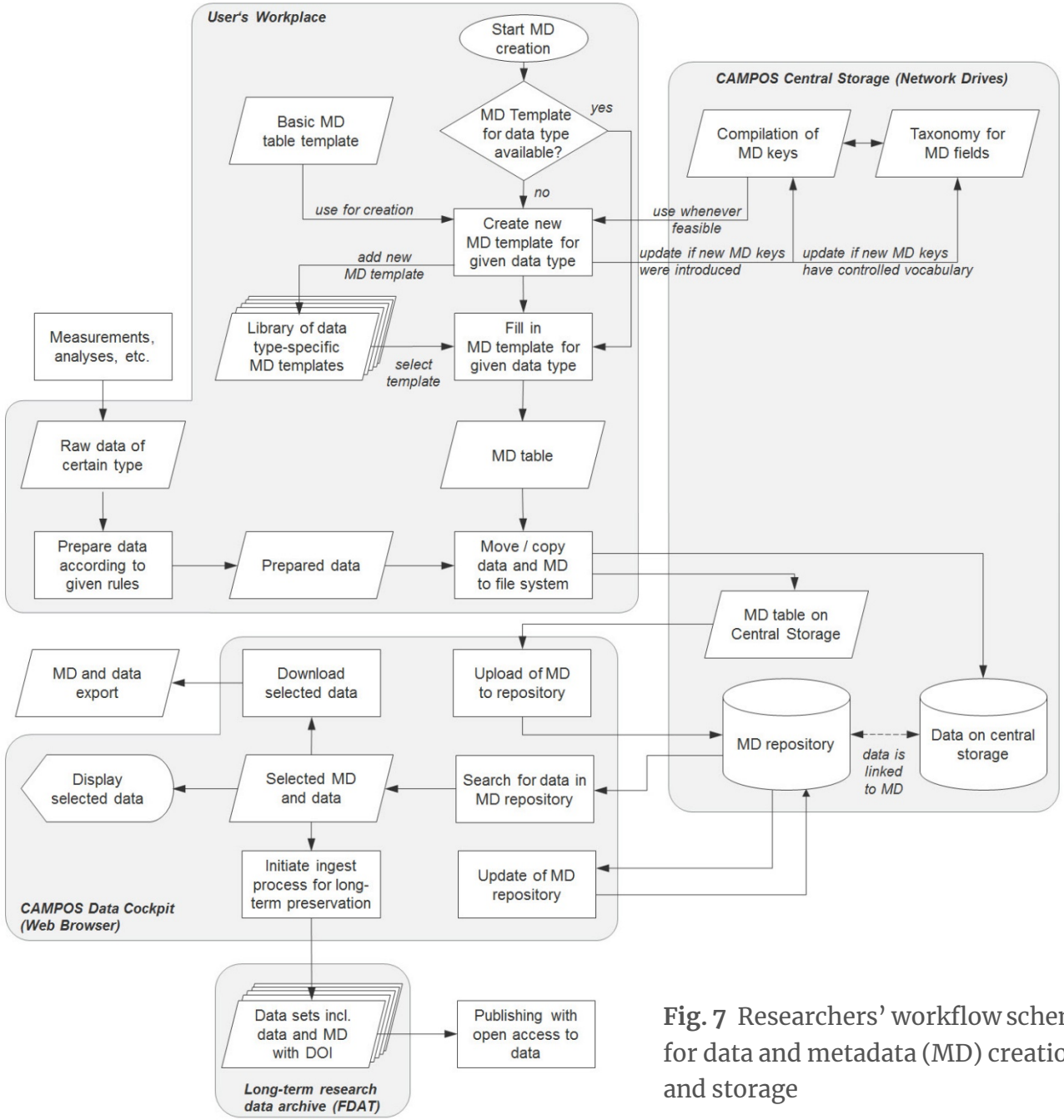


Fig. 7 Researchers' workflow scheme for data and metadata (MD) creation and storage

Time-line of implementing the data management framework

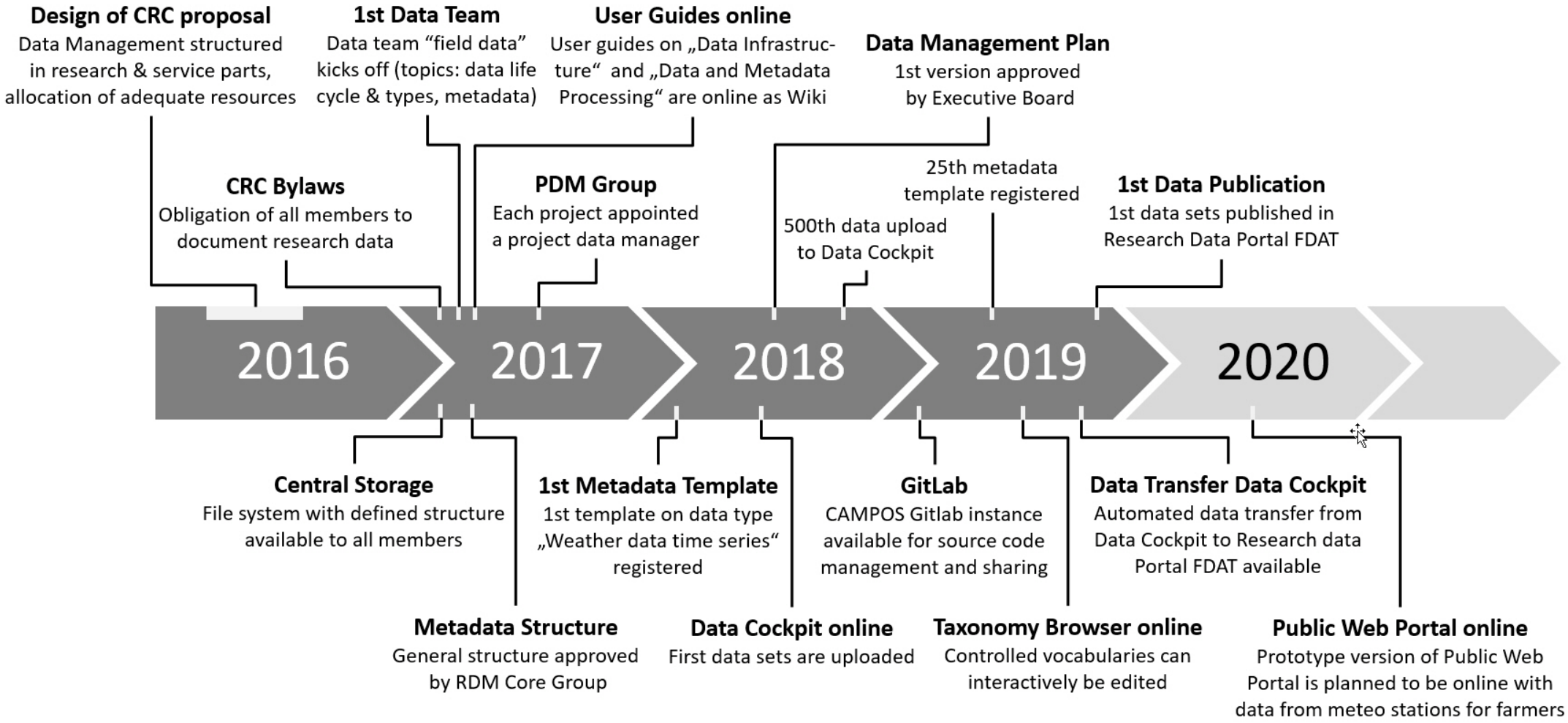


Fig. 8 Time line of the development and implementation of the CAMPOS data management framework (Source: Finkel et al. 2020)

Further work

Further streamlining of the metadata creation process

through electronic laboratory and field notebooks and tools for automated metadata creation (such as the generation of series of metadata of similar type)

Enabling automated generation of user interfaces and database definitions

using a metadescription-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 allow a generic implementation 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 would render own metadata tooling solutions unnecessary.

Tools and workflows for data visualization, exploitation, and exchange

Management of modelling data

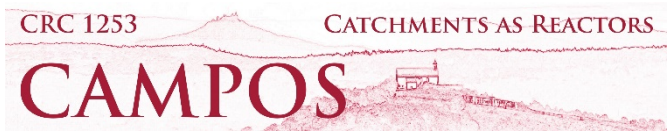
Acknowledgement, address and links

This work was supported by the Collaborative Research Center 1253 CAMPOS, funded by the German Research Foundation (DFG, Grant Agreement SFB 1253/1 2017).

Dr. Michael Finkel

Eberhard Karls Universiy Tübingen
Faculty of Science
Center for Applied Geosciences
Hölderlinstrasse 12 · 72074 Tübingen
Fon +49 (0)7071 2975022 · Fax +49 (0)7071 295059
michael.finkel@uni-tuebingen.de

Homepage of CRC CAMPOS



This presentation material file is distributed under the
Creative Commons Attribution 4.0 International License (CC BY 4.0)