Terrestrial Environmental Observatories (TERENO)

Data Management Plan

This Project Communication and Data Management Plan was produced by the TERENO Coordination Team Data Management (CT-DM) at the outset of the TERENO project, and has been approved by the TERENO Scientific Steering Committee (SSC). The plan is intended to help research teams and the contributing institutions to think about their data management organization, activities and responsibilities, in particular communication and dissemination activities, engagement with stakeholders, and data management needs and responsibilities.

The plan aims to describe the approaches necessary to implement a TERENO data management infrastructure. It concerns the management and exchange of data and data products to be created by the individual institutions contributing to TERENO. The plan acts as an agreement between data users and data providers/creators in the framework of TERENO, on how to use, store, disseminate and publish data generated by the project.

The target audience for this plan is all project members from the contributing institutions as well as partner organizations providing and using data and data products.

The Plan also provides a basis for the overall planning and development of the central TERENO data management by:

- Providing information that can be used to co-ordinate communication activities and stakeholder relations across TERENO.
- Highlighting data management and custody issues so that data is managed in a way that meets the requirements of the TERENO project management, and enables the project to respond to common data needs.
- Providing a basis for quality assurance within the project.
- Providing a basis from which TERENO project partners and the project management can report and monitor project and overall TERENO project progress.

Document version

V 1.0

Document date

2014-01-23

Authors:

- Ralf Kunkel, Research Centre Jülich, 52425 Jülich, Germany, E-Mail: r.kunkel@fz-juelich.de
- Martin Abbrent, Helmholtz-Centre for Environmental Research – UFZ, 04318 Leipzig, Germany, E-Mail: martin.abbrent@ufz.de
- Anusuriya Devaraju, Research Centre Jülich, 52425 Jülich, Germany, E-Mail: a.devaraju@fz-juelich.de
- Jan Friesen, Helmholtz-Centre for Environmental Research – UFZ, 04318 Leipzig, Germany, E-Mail: janfriesen@ufz.de
- Rainer Gasche, Karlsruhe Institute of Technology (KIT), Institute of Meteorology and Climate Research – Atmospheric Environmental Research (IMK-IFU), 82467 Garmisch-Partenkirchen, Germany, E-Mail: rainer.gasche@kit.edu
- Jens Klump, former GFZ German Research Centre for Geosciences, 14473 Potsdam, Germany, now CSIRO ARRC, 26 Dick Perry Avenue, Kensington, WA 6151, Australia, E-Mail: Jens.Klump@csiro.au
- Frank Neidl, Karlsruhe Institute of Technology (KIT), Institute of Meteorology and Climate Research – Atmospheric Environmental Research (IMK-IFU), 82467 Garmisch-Partenkirchen, Germany, E-Mail: frank.neidl@kit.edu
- Karsten Rink, Helmholtz-Centre for Environmental Research – UFZ, 04318 Leipzig, Germany, E-Mail: karsten.rink@ufz.de
- Thomas Schnicke, Helmholtz-Centre for Environmental Research – UFZ, 04318 Leipzig, Germany, E-Mail: thomas.schnicke@ufz.de
- Matthias Schroeder, GFZ German Research Centre for Environmental, 14473 Potsdam, Germany, E-Mail: matthias.schroeder@gfz-potsdam.de
- Jürgen Sorg, FZJ, Research Centre Jülich, 52425 Jülich, Germany, E-Mail: j.sorg@fz-juelich.de
- Vivien Stender, GFZ German Research Centre for Geosciences, 14473 Potsdam, Germany, E-Mail: vivien.stender@gfz-potsdam.de
- Ute Wollschläger, Helmholtz-Centre for Environmental Research – UFZ, 04318 Leipzig, Germany, E-Mail: ute.wollschlaeger@ufz.de
- Steffen Zacharias, Helmholtz-Centre for Environmental Research – UFZ, 04318 Leipzig, Germany, E-Mail: steffen.zacharias@ufz.de

Content Scheme:

- Data Management Planning Tool "DMP online" (https://dmponline.dcc.ac.uk/) by the Digital Curation Centre (http://www.dcc.ac.uk/)
  Version: 06
  Date: 2013-06-01

Document history:

2013-10-31	First draft 0.1 (Ralf Kunkel)
2013-12-06	Revision according to remarks made at CT DM meeting,
2013-12-18	Additional revisions according to the remarks of authors
2014-01-23	Final version to be approved by the Scientific Steering Committee
2014-05-28	Approval by the Scientific Steering Committee

Contents

1 Glossary

CSW
Catalog Service for the Web: Standard of the Open Geospatial Consortium, which defines common interfaces to discover, browse, and query metadata about data, services, and other potential resources.

CMC
Central Metadata Catalogue, which stores all metadata managed by TEODOOR.

CT-DM
Coordination Team Data Management: this group consists of the data managers responsible for conceptualization and implementation of the data management infrastructure in TERENO. It is responsible for the overall coordination, to build up and operate the TERENO data management infrastructure and to organize and carry out user trainings.

DMI
Data Management Infrastructure

IDM
Institutional Data Manager: nominated person responsible for data management issues of a particular TERENO institution.

OGC
The Open Geospatial Consortium (OGC), an international voluntary consensus standards organization, established in 1994. In the OGC, more than 400 commercial, governmental, nonprofit and research organizations worldwide collaborate in a consensus process encouraging development and implementation of open standards for geospatial content and services, GIS data processing and data sharing.

SDI
Spatial Data Infrastructure: Data infrastructure implementing a framework of geographic data, metadata, users and tools that are interactively connected in order to use spatial data in an efficient and flexible way.

SOS
Sensor Observation Service: Web service to query real-time sensor data and sensor data time series and is part of the OGC Sensor Web Enablement framework.

SWE
OGC's Sensor Web Enablement framework defines a suite of web service interfaces and communication protocols abstracting from the heterogeneity of sensor (network) communication.

UID
Unique Identifier: any identifier, which is guaranteed to be unique among all identifiers, used for those objects and for a specific purpose.

TERENO
TERrestrial ENvironmental Observatories

TERENO-portal
Central web-portal of the TERENO initiative, accessible through http://www.tereno.net. This portal allows querying, visualizing and accessing data and metadata from the different observatories in a standardized way and thus acts as a database node providing scientists and stakeholder with reliable and well accessible data and data products.

TEODOOR	TErenO Data Online repositORy: a spatial data infrastructure created to manage and publish the data from the TERENO project.
WCS	The Open Geospatial Consortium Web Coverage Service Interface Standard defines Web-based retrieval of raster data (coverage)
WFS	Web Feature Service: A standard interface from the Open Geospatial Consortium that allows requests for geographical features across the web using platform-independent calls.
WMS	Web Map Service: Standard protocol for serving geo-referenced map images over the Internet that are generated by a map server using data from a GIS database. The specification was developed and first published by the Open Geospatial Consortium in 1999.

2 Basic Project Information

TERENO (TERrestrial ENvironmental Observatories) is an initiative funded by the large research infrastructure program of the Helmholtz Association within the research area "Earth and Environment". The main goal of TERENO is to create observation platforms on the basis of an interdisciplinary and long-term aimed research program to investigate the consequences of Global Change for terrestrial ecosystems and the socioeconomic implications. Special attention is paid to bridge the different temporal and spatial scales by covering the investigated systems – land surface, biosphere, lower atmosphere and anthroposphere from the local to the regional scale and from direct observable time periods (years) to long time periods (centennials) derived from geo-archives. TERENO provides time series of system variables (e.g. precipitation, runoff, groundwater level, soil moisture, water vapor and trace gases fluxes) for the analysis and prognosis of global change consequences using integrated model systems, which will be used to derive efficient prevention, mitigation and adaptation strategies.

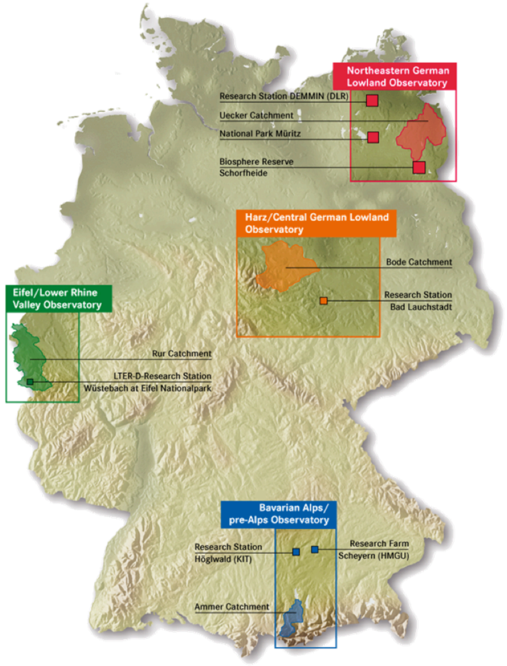
Northeastern German Lowland Observatory
Research Station DEMMIN (DLR)
Uecker Catchment
National Park Müritz
Biosphere Reserve Schorfheide
Harz/Central German Lowland Observatory
Bode Catchment
Research Station Bad Lauchstadt
Eifel/Lower Rhine Valley Observatory
Rur Catchment
LTER-D-Research Station Wüstebach at Eifel Nationalpark
Bavarian Alps/ pre-Alps Observatory
Research Station Höglwald (KIT)
Research Farm Scheyern (HMGU)
Ammer Catchment

Figure 1: Location of the TERENO observatories in Germany (Zacharias *et al.*, 2011).

Four terrestrial observatories located across Germany (see figure 1) have been selected to be representative to indicate the long-term effects of climate and land use changes. The observatories are operated by different Helmholtz Centers. Installation of equipment started in 2007 and is finished in 2013. Data collection, however, will be performed for at least 15 years.

3 Capture Methods and Management of Research Data

3.1 Overview of the TERENO Data Infrastructure Setup

During instrumentation of the four terrestrial observatories local data infrastructures have been implemented by the individual Helmholtz Centers, in which the TERENO data were integrated successively. Even if the four observatories have some differences in the main focus of research, the collected data are very similar and can be distinguished into three types:

- File based non-geospatial data (documents, reports etc.)
- Campaign data, laboratory analyses
- Geospatial data without significant temporal variations (soil maps, geological maps etc.)
- Time series data.

By following the concepts of a Spatial Data Infrastructure (SDI), where spatial (and non-spatial) data, metadata, users and tools are interactively connected, the decentralized data infrastructure TEODOOR (TEreno Online Data repOsitORry) is being established within the framework of TERENO. This decentralized setup aims at interconnecting regional data and metadata infrastructures established under the "umbrella" of TERENO A central "umbrella" portal application allows to query, visualize and access all data and metadata (depending on the data owners) from the contributing institutions in a standardized way and acts in this way as a database node providing scientists and stakeholder with reliable and well accessible data and data products.

Data publication and exchange is facilitated predominantly through web services as standardized by the International Standardization Organization (ISO) and the Open Geospatial Consortium (OGC, http://www.opengeospatial.org/), operated from the data providers. TEODOOR contains the following basic features (see figure 2):

1. A portal application, which allows one to query, find and access data from local TERENO or external data infrastructures according to the data policy.

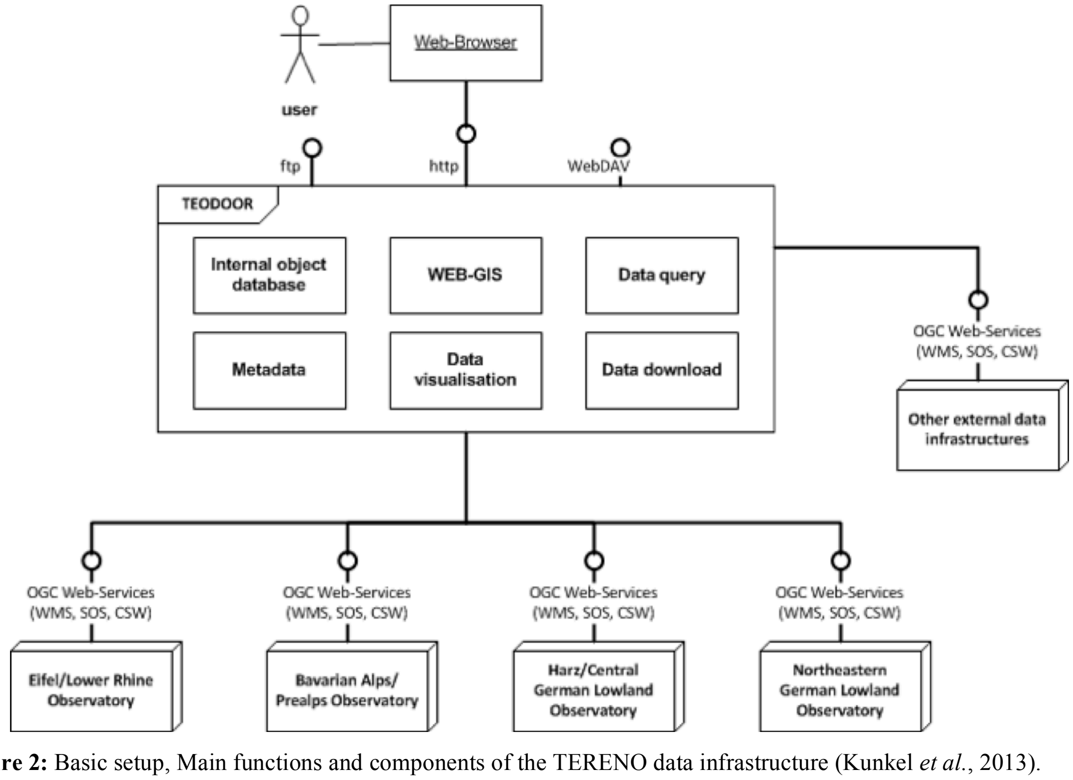
Web-Browser
user
ftp
http
WebDAV
TEODOOR
Internal object database
WEB-GIS
Data query
Metadata
Data visualisation
Data download
OGC Web-Services (WMS, SOS, CSW)
Other external data infrastructures
OGC Web-Services (WMS, SOS, CSW)
Eifel/Lower Rhine Observatory
OGC Web-Services (WMS, SOS, CSW)
Bavarian Alps/ Prealps Observatory
OGC Web-Services (WMS, SOS, CSW)
Harz/Central German Lowland Observatory
OGC Web-Services (WMS, SOS, CSW)
Northeastern German Lowland Observatory

Figure 2: Basic setup, Main functions and components of the TERENO data infrastructure (Kunkel *et al.*, 2013).

2. Local data infrastructures hosted by the individual TERENO institutions
3. Defined interfaces for data exchange between the portal application and local data infrastructures provided by standardized, OGC-conformal web-services operated at FZJ and, as far as existent, in regional data infrastructures maintained by partners. The following web-service specifications is used:
   - OGC Sensor Web Enablement (SWE) standards and interfaces are used to provide time series data from the TERENO observation networks together with contextual details such as instruments used, intended application, valid time, contact details to responsible persons, security level, geographic location.
   - OGC compliant Web Map Services (WMS), Web Feature Services (WFS) and Web Coverage Services (WCS) are implemented by using a web-map server (Geoserver, ESRI GisServer) to provide map layers generated from vector data or raster data to the users. Non-OGC compliant web services may be implemented to provide additional functionalities not covered by the OGC standards and interfaces specifications. Common data file download services (based on http, FTP, etc.) are connected to the aforementioned services where appropriate.
   - Metadata are published by OGC compliant Web Catalogue Services based on the ISO 19115/19139 metadata standards. Beneath Catalogue services operated by the individual partner institutions, a Central Metadata Catalogue (CMC) as a common entry point to TERENO metadata is be implemented at Research Centre Jülich.

3.2 Third party and external data

Existing input data sets of different types (observations/raw data, derived data, etc.), required for the work within TERENO are provided by external sources. Data sources in this respect are collected data or data produced by previous research projects, or provided by governmental institutions, non-governmental organizations and programs as well as by international programs in the TERENO regions. Access to the data of interest stored in the data repositories that are maintained by the individual TERENO partner institutions is regulated by the TERENO Data Policy document.

3.3 New Data

One of the most important targets of TERENO is the generation of new data to fill gaps in knowledge on regional phenomena, and to feed a variety of models and tools for scenario development, which serve as the scientific basis for policy advice to national/regional/local stakeholders. Within TERENO new data are captured or created by

- Time series data, e.g. hydrological and meteorological observation data documenting the long-term climate change, gathered by regional observation networks,
- Biodiversity data,
- Socio-economic data,
- Remotely sensed images in different spatial and temporal resolutions,

- Data collection by individual researchers from external sources, own investigations and processed data.

3.4 Data integration and standardization

In order to ensure compatibility in data integration by several applications/models, detailed standards and regulations for particular data groups, which are not covered by common standardization rules, are specified in cooperation with the scientific experts in their fields. Integration between the data being gathered in the project and pre-existing data sources are made on several levels, as far as appropriate:

- Observable property level: where needed, transformation of units and recalculations of values
- Format level: reformatting to standardized formats (see section 3.5)
- Geographic reference system level: transformation to standardized coordinate systems
- Documentation level: use of standardized metadata profiles
- Temporal and spatial scales.

The CT Data Management provides reference tables to the TERENO community with mandatory standard vocabularies to be used within TEODOOR.

3.5 Data Formats

Data storage and data formats highly depend on the types of data, which can be very heterogeneous. Basically, data of interest in TERENO can be classified as follows:

- Structured (quantitative) data:
  - Time series data provided by online in-situ or remote sensors, to be stored in relational databases
  - Tabular (statistical) data derived from field campaigns, surveys or generated by simulations, to be stored in files or relational databases
- Unstructured (qualitative) data such as descriptive documents, audio- and video data, pictures, stored in files
- Geo-referenced data stored in files or relational geo-databases
  - Vector-based geospatial data
  - Raster based geospatial data

Except for unstructured information, all data are being interpreted by computer programs to make it understandable and are - by its very nature - software dependent. Data are thus endangered by the obsolescence of the hardware and software environment on which access to data depends. The best option to warrant interoperability of data between varieties of applications and over the long term is to convert data in non-proprietary and standardized open-formats. Nevertheless, archiving of data in proprietary or quasi-proprietary formats (e.g. MS Word, MS Excel) can be required in cases where meaningful content would get lost when storing in open formats, e.g. formulas and diagrams in MS Excel tables when exporting in comma-separated values (ASCII/CSV) files. In order to save storage space on the server's hard-drives, uploaded data may be stored as open compressed data file packages (e.g. .zip, .7z). Local data managers may be contacted if data from non-proprietary or not supported file formats need to be converted into supported formats.

The following data file formats for data are supported within TEODOOR (the preferred data type is typed in bold letters):

Type of data	**Supported formats for exchange and reuse**
Qualitative data Textual documents	• Rich text format (rtf), • Open Document Text (.odt), • Hypertext markup language (.htm), • Plain Text (.txt), • eXtensible Mark-up Language (XML) text according to an appropriate Document Type Definition (DTD) or schema (.xml), • **Portable Document Format (.pdf),** Widely used office formats (e.g. MS Word)
Quantitative tabular data (with or without column labeling, variables names and metadata)	• **Comma Separated Values (.csv),** • Tab-delimited File (.tab), • Open Document Spreadsheets (.ods), • Extensible Markup Language (.xml), • NetCDF (.netcdf), • Hierarchical Data Format (.hdf), • Microsoft Excel (.xls, .xlsx), DBase (.dbf)
Digital image data	• Bitmap (.bmp) • **Joint Photographic Experts Group (.jpg, .jpeg)** • **Portable Network Graphics (.png)** • **Tagged Image File (.tif, .tiff)**
Georeferenced vector data	• **ESRI shape file (essential - .shp, .shx, .dbf, optional - .prj, .sbx, .sbn)** • Geographic Markup Language (.gml) • Scalable Vector Graphics (.svg) • Keyhole Mark-up Language (KML) • MapInfo Interchange Format (.mif)
Georeferenced raster data	• **GeoTIFF (Tagged Image File Format/.tif, .tiff),** • **NetCDF (.netcdf)** • **Esri ASCII or binary Grid (.asc, .flt)** • Hierarchical Data Format (.hdf)
Digital video data	• **MPEG-4 (.mp4),** • Motion JPEG 2000 (.mj2)
Compressed file formats	• **Zip-format (.zip)** • 7-Zip-format (.7z)

Due to the extensive data volumes and for reasons of accessibility, time series data from in-situ and remote sensors are not stored in simple file based formats but in relational database systems. For data storage in the database and registration of the sensor metadata, an underlying data model for time series data is used, which stores observation data along with sufficient metadata to provide traceable heritage from raw measurements to usable information.

3.6 Quality Control

The level of data quality, following the common definition of data quality i.e. "The state of completeness, validity, consistency, timeliness and accuracy that makes data appropriate for a specific use"[1], is dependent on the real world phenomena represented by the data and data products. To carry out data quality assessments, expert knowledge on the field of data generation is demanded. Therefore, the responsibility for data quality assessments and assurance measures is by the nominated responsible persons working on the data. These persons are also responsible to describe the data quality assessment measures and methods and to provide an evaluation on the trustworthiness of the data within the Metadata Catalogues.

The responsible persons must attribute all TERENO data with respect to its data quality and data processing levels. This is to be done by a combination of a data qualifier and a processing status flag assigned to each data set. These quality control information are either stored in the metadata for entire data sets (geodata or other file based data) or on the level of individual observations (for times series data).

The data qualifiers (aka. quality flags) provide qualifying information that can note validity information about the data (like “visually checked”) or anything unusual or problematic about individual observations (such as, e.g., "holding time for analysis exceeded" or "incomplete or inexact daily total"). Due to the diverse nature of TERENO observations, a common set of quality flags is required that can be used by different sensing applications. Following (UNESCO, 2013), a two-level flag scheme is used. The first level defines the generic data quality flags, while the second level complements the first level by providing the justification for the quality flags based on validation tests and data processing history. Whereas the generic flags are fixed, the second-level flags are specified by the domain experts and can be extended as necessary. The characterization of these flags can be sensor-specific or property-specific, depending on the application. The following table shows the generic flags and examples of specific flags.

Generic flag	**Examples of specific flags**
Unevaluated	-
Ok	Good quality, moderate quality, passed auto check
Bad	Below minimum/ above maximum value, isolated spike, defective sensor
Suspicious	Same as “bad”
Gap filled	Interpolated, extrapolated
Missing	-

[1] http://en.wikipedia.org/wiki/Data_quality

A list of data qualifiers were elaborated by the CT Data Management in collaboration with the responsible persons, published as controlled common vocabulary by TEODOOR, and are continuously be revised on demand.

3.7 Data processing levels

All data created, collected and used within the TERENO project and being regarded as valuable for verifying scientific findings, reusing in further research and integrating in policy advice statements, are stored in TEODOOR or in to TEODOOR connected data infrastructures. However, in order to avoid overflows in the data repositories and the accumulation of unnecessary and unneeded data versions (see section 3.9), responsible persons should carefully evaluate if a dataset should be shared by integration into the databases or not.

The processing status characterizes the general processing steps that the data have been subjected to. Depending on the parameters being controlled, several categories of data are defined:

a) **Raw data** is defined as unprocessed or preliminary pre-processed data and data products that have not undergone quality control. Depending on the data type and data transmission system, raw data may be available within seconds or minutes after real-time.

b) **Quality-controlled data** have passed quality control procedures such as routine estimation of timing and sensor calibration or statistical/visual inspection and identification of obvious errors.

c) **Derived Products** have to be derived from quality-controlled data. They require scientific and technical interpretation and may include multiple-sensor data, researcher (PI) driven analysis and interpretation, model-based interpretation using other data and/or strong prior assumptions.

3.8 User groups

With the intention to describe phases in creating scientific and publishable findings, the state of data processing can be related to specific research domains. Research Domains define the group of users with whom data are shared (Treloar & Harboe-Ree, 2008). The listing below serves as an orientation guide for the decision as to whether data should be up-loaded into TEODOOR or not. If in doubt, the Coordination Team Data Management or the project coordinators should be consulted.

a) **Private Research Domain** involves members of an individual research team. Data used in this domain usually represent raw, preliminary or intermediate rather than final research findings. Data management and sharing occurs within the team, and is pre-defined as members only.

   → Therefore, data do not have to, but may be integrated in TEODOOR. If data are integrated into TEODOOR, it has to comply with the supported formats and becomes accessible with restrictions, such as for instance for project members only. Storage time of data is restricted to 12 months, after which data have to be transferred into the shared or public domain or being deleted.

b) **Shared Research Domain** involves researchers collaborating within TERENO, often across institutions. Data shared mostly represent intermediate research results, which are not the subject of a scientific publication as yet or external data, whose usage is granted only to TERENO members. Data should have to be undergone quality assurance procedures.

→ Therefore, data have to be integrated in TEODOOR in a supported format, but the data access is restricted to TERENO members (mapped as “user groups” in the system).

c) **Public Domain** involves the public sphere. Data had to be undergone quality assurance procedures and may be used in or referenced by publications.

→ Therefore, data have to be integrated in TEODOOR in a supported format, and are made accessible by the scientific community, decision makers, politicians and the interested, non-commercial public.

3.9 Versioning

Within a data processing chain several versions of a data set may be created. A version can represent a status in the processing chain, or a small modification made to the data due to other reasons. Versioning might also occur in scenarios being result of modeling.

Versions can be marked simply by adding a version number to the file-, database-name, database-record or to a data-identifier, e.g. <file/database/database-record-name><version-number>, respectively, by adding a time stamp. In addition, parent/child relationships may be included in the Metadata Catalogues to document several states in the processing chain as versions.

Where required server-client based version control software is used, e.g. Apache Subversion (SVN). Subversion management software is commonly used in software development environments, where several developers work on the software code locally, and code extensions or modifications are merged to a software code repository on a server. SVN clients are connected to the server by common data transfer protocols as like FTP, SFTP or SCP.

The possible usefulness of the above, and implementation for other applications e.g. climate modeling in the framework of TEODOOR is discussed with the respective specialists in the individual TERENO Coordination Teams on demand.

3.10 Data Documentation and Metadata

Data documentation through metadata is an essential requirement for the search, assessment and evaluation of data within a data infrastructure. Moreover, the task of a data infrastructure to disseminate and to promote the usage of the data usually fails if the data are not properly described and, as a consequence, users are not able to judge whether the data content is relevant and the data quality is sufficient to their specific needs. Therefore, all data have to be accompanied by significant metadata.

In TERENO, these metadata are managed and published and, in case of file-based data manually provided to TEODOOR, be created by using the tools provided by Metadata Catalogue applications operated by each TERENO institution. Alternatively, a Central Metadata Catalogue (CMC) is hosted at FZJ and can be used by the responsible persons to insert and manage metadata sets. A metadata set consists of several elements to describe the data, e.g. author, abstract, lineage, data quality, etc. In order to exchange metadata with other metadata catalogues or comparable applications, international approved standards and protocols are used (see section 3.10.1).

3.10.1 Metadata Standards

The metadata standard for file based data and geodata to be used is the ISO 19115 "Geographic Information - Metadata" standard, in xml-encoding following ISO 19139. These standards are widely established within distributed spatial data infrastructures, where geo-referenced data represent a large proportion of the data stock. The set of metadata elements in the ISO standards is capable to cover most data-types, data-topics and data–processes, even for data without coordinate-based spatial reference. If necessary, the ISO standard 19110 to describe feature types (e.g. in tabular data on attribute level) were used. For digital documents (qualitative) and any other type of qualitative digital web-content, Dublin Core is used as metadata standard.

Time series data gained from the observation network is distributed by web-services, which provides the data together with basic metadata. The metadata scheme provided by these services corresponds to the SensorML specification based on the Sensor Web Enablement (2.0) initiative of the OGC.

3.10.2 Metadata Profiles

Every user group is being provided with a range of metadata profiles/templates tailored to the requirements of specific data groups (e.g. time series data, remote sensing data, survey data, statistical data) under their custody. The profiles are subsets of the basic standard ISO 19115, which consists of around 400 metadata elements, and which help the users to work more economically with the used metadata editors. The used profiles have to comply with the EU-INSPIRE directive, were developed by the CT DM, together with responsible persons from the TERENO research groups, and implemented in the Metadata Catalogues as far as is suitable.

3.10.3 Metadata Creation and Management

Creation and capturing of metadata depends on the type and origin of the research data. Some applications (e.g. ESRI ArcGIS) are capable to create and export metadata into a standard, which is compliant to the TERENO metadata standards. In general, however, metadata for file-based data are created online through editing tools within the individual Catalogue Service applications hosted by the individual observatories or by those of the CMC. The life cycle of metadata, from creation and content controlling of a metadata set to its publication in the web, is managed by dedicated user profiles in the system. The members of the CT Data management train the local users to use the individual metadata-editing tools. The process of establishing these metadata creating and data sharing workflows is led and supported by the CT Data Management as well. In the creation and management of metadata the following rules apply:

a) Every data set stored or published by TEODOOR must be accompanied with metadata, compliant to the defined metadata standards and profiles.

b) It is the responsibility of the responsible persons providing data and proper metadata to TEODOOR.

c) For every TERENO institution at least one responsible person is nominated (see section Annex 1). This person is in charge to control the upload and the distribution of data the content and the compliance to the standards of the metadata contributed by colleagues in the particular institution.

d) The metadata must contain either a URL to the online resource for accessing the data directly or at least the contact details where the data can be retrieved when stored locally.

e) The owners can modify metadata on any occasion, but it must contain a basic set of metadata elements – mostly part of by a dedicated metadata profile (see section 3.10.2) - to be classified as “completed”.

f) Data are rejected from being imported into TEODOOR if proper and completed metadata are not delivered within one month after uploading the data.

In addition, controlled vocabularies such as keyword-thesauri, reference tables, fixed categories etc. were implemented in coordination with the users, to support the easy retrieval of data by filtering the metadata. The General Multilingual Environmental Thesaurus (GEMET) is the first choice for thematic thesaurus implementation, together with a thesaurus for spatial locations. Furthermore, for example, more scientific discipline oriented thesauri, may be included in the catalogue if required by the users.

3.11 Geo-Referencing

In order to create maps by overlaying geospatial layers generated from geodata of different sources it is important that the layers share the same geographic coordinate systems (GCS) and projections. Specific requirements on the accuracy of spatial localization may demand the use of regional or local coordinate systems and projections. However, when submitting geospatial layers to TEODOOR the geodata should be referenced by the World Geodetic System 1984 (WGS 1984, EPSG: 4326). Therefore it might be required to transform the geodata from other geodetic reference systems, e.g. UTM or Gauss-Krueger into WGS 1984, before uploading into TEODOOR or publishing the geodata as Web Map Service layers. The original geodetic reference system can be included in the upload.

4 Data Rights Management: Privacy and Intellectual Property

Data sharing and re-use is subject to legal regulations. Therefore, the data rights management is of crucial importance in the framework of data management, and of high priority in each and every Data Management Plan. This section in the Data Management Plan was developed by the CT-DM in coordination with the responsible bodies in the TERENO project consortium, and with the help of legal experts. A separate Data Policy document is attached to this Data Management Plan.

5 Data Sharing, data provision, citation rules

Project data hosted by TEODOOR have to be accessible to project partners, and respectively to the public, as soon as this is demanded by the research teams and/or the site coordinators and/or the Scientific Steering Committee. Details of data sharing, data provision and citation regulations are specified in the TERENO Data Policy document.

6 Short-Term Storage and Data Management

6.1 Storage Media and Data Transfer

Data from partners are stored in their own data infrastructures that are linked to TEODOOR by standardized web services (see section 3.1). Independently of the actual data storage devices in the TERENO data infrastructure, a registration of data sets in the metadata base together with distribution information giving instructions where to ask for data has to be made at least.

6.2 Backup

Loss of data has to be prevented by data backup. Data backup is accomplished and under the responsibility of each TERENO institution performing regular data and system (virtual machines) backups on hard disk drives and regular data archiving. Backup frequencies depend on the data modification rate but have to least once per day. Archive of data is performed if necessary. The responsibility for the implementation and supervision of proper data backup and archive strategies at an institution is with the Institutional Data Managers (IDM, see section 8.1) of a particular TERENO institution.

6.3 Security

During the project's lifetime access restrictions and data security is ensured by using common IT components such as firewalls, virus-scanners, secure connections. Access to data provided by internal and external web services may be restricted to take place via the TERENO portal. Based on data policy agreements, users have to to be registered in the data infrastructure. Every user accessing TERENO data ought to accept and sign an agreement in which the license conditions are clearly specified. Access control is implemented on a per dataset basis in the metadata catalogue or through network access control, e.g. when accessing the server via FTP, SSH or other web-protocols.

Based on user profiles and user groups, the TERENO-portal gives dedicated access and activity permissions on the system to a user. The Coordination Team Data Management in cooperation with the Institutional Data Managers manages access to data.

7 Deposit and Long-Term Preservation

A long-term strategy for maintaining, administering and archiving the data is to be ensured by the long-term storage of data at the individual TERENO institutions.

7.1 Long-Term Specifics

Data have to be kept for at least 15 years under the responsibility of each TERENO institution owning the data. Life cycles of data are handled by following the demands expressed by the involved investigator and stakeholders. Plans for archiving and preservation strategies (by considering archived data as not immediately accessible data), are created by the CT-DM in time.

Managing datasets including sensitive data over the longer term will be accomplished taking into consideration national laws.

Data transformation by reformatting to open formats or by transformation to an appropriate geographic reference system may be performed, if necessary.

7.2 Metadata and Documentation for Long-Term Preservation

Datasets are linked by internal entries with the UIDs in the metadata base. In addition, xml files (ISO 19139 standard) are to be downloaded automatically together with the data files when a user accesses the data in TEODOOR. This is done by the data providers/creators by editing metadata online on the TERENO data portal, or by using appropriate tools extracting meta information from data stored on local drives in a format (19139 xml schema), which can be uploaded to the central metadata catalogue.

Published materials and/or outcomes, preferentially linked to the used data are included in the data infrastructures. This is done by persistent URLs (composed of server hostname and metadata UID) to the related entry in links in the metadata catalogue, respectively by Digital Object Identifiers (e.g. from the DOI-System, see http://www.doi.org/) as far as is introduced in TERENO.

7.3 Longer-Term Stewardship

Responsibility over time for decisions about the data once the original personnel have left is with the C T Data Management together with TERENO Scientific Steering Committee, and based upon data policy statements agreed with the respective institutional bodies of TERENO.

8 Resourcing

8.1 Organizational Roles and Responsibilities for Data Management

During the TERENO project funding time period, coordination of data management at different levels of responsibility is established in the following manner:

a) Coordination Team Data Management (CT-DM): The CT DM consists of the data managers responsible for conceptualization and implementation of the data management infrastructure in TERENO. This group is responsible for the overall coordination, to build up and operate the TERENO data management infrastructure and to organize and carry out user trainings. Members of the group are listed in Annex A 1.1.

b) Institutional Data Managers (IDM): The institutional data manager is a nominated person responsible for data management issues of a particular TERENO institution. The IDM are responsible for controlling, processing, preparation for sharing (e.g. by transformations), quality assessment and control of data to be integrated in TEODOOR, and for provision of metadata with respect to the TERENO data management policies. The CT-DM works closely together with the IDM by supporting and training them in data management tasks within their responsibility e.g. the submission of metadata and data to the central databases, and organizes further processing as far as this falls under their responsibility. Further processing, aside of data sharing, could be the creation of portrayal and other visualization services. The group meets and communicates on demand.

Funding of data management within the project lifetime is part of the proposed project budget. Longer term funding after the project’s end is not yet specified.

9 Adherence and Review

The degree of adherence to the plan is observed in the daily work of the Institutional Data Managers and checked by the CT Data Management. If necessary, this plan will be revised by the CT Data Management.

10 Statement of Agreement

This data management plan has been agreed upon by the TERENO Scientific steering Committee on 2014-05-28.

11 References

Kunkel, R., Sorg, J., Eckardt, R., Kolditz, O., Rink, K., Vereecken, H. (2013): TEODOOR: a distributed geodata infrastructure for terrestrial observation data. Environmental Earth Sciences (2): 507-521, doi: 10.1007/s12665-013-2370-7.

Treloar, Harboe-Ree, 2008. http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf.

UNESCO, I.o., 2013. Ocean Data Standards: Recommendation for a Quality Flag Scheme for the Exchange of Oceanographic and Marine Meteorological Data, IOC Manuals and guides No. 54 - Volume 3. http://www.iode.org/components/com_oe/oe.php?task=download&id=20559&version=1.0&lang=1&format=1.

Zacharias, S., Bogena, H., Samaniego, L., Mauder, M., Fuss, R., Puetz, T., Frenzel, M., Schwank, M., Baessler, C., Butterbach-Bahl, K., Bens, O., Borg, E., Brauer, A., Dietrich, P., Hajnsek, I., Helle, G., Kiese, R., Kunstmann, H., Klotz, S., Munch, J.C., Papen, H., Priesack, E., Schmid, H.P., Steinbrecher, R., Rosenbaum, U., Teutsch, G., Vereecken, H. (2011): A Network of Terrestrial Environmental Observatories in Germany. Vadose Zone Journal (3): 955-973, doi: 10.2136/vzj2010.0139.

Annex 1 Nominated data managers

A 1.1 Coordination Team Data Management Members

- Ralf Kunkel (head of group)
  Research Center Jülich
  Institute for Bio and Geosciences – Agrosphere (IBG-3)
  52425 Jülich, Germany
  Phone: +49-2461-61-3262
  E-Mail: r.kunkel@fz-juelich.de
- Martin Abbrent
  Helmholtz-Centre for Environmental Research – UFZ
  Department Environmental Informatics
  Permoserstraße 15
  04318 Leipzig, Germany
  Phone: +49-341-235-1903
  E-Mail: martin.abbrent@ufz.de
- Anusuriya Devaraju
  Research Center Jülich
  Institute for Bio and Geosciences – Agrosphere (IBG-3)
  52425 Jülich, Germany
  Phone: +49-2461-61- 1652
  E-Mail: a.devaraju@fz-juelich.de
- Mark Frenzel
  Helmholtz-Centre for Environmental Research – UFZ
  Department Community Ecology
  Theodor-Lieser-Str. 4
  06120 Halle (IMK-IFU),
  Phone: +49-345-558-5304
  E-Mail: mark.frenzel@ufz.de
- Rainer Gasche
  Karlsruhe Institute of Technology (KIT),
  Institute of Meteorology and Climate Research – Atmospheric Environmental Research (IMK-IFU),
  82467 Garmisch-Partenkirchen, Germany
  Phone: +49-8821-183-132
  E-Mail: rainer.gasche@kit.edu
- Jens Klump
  GFZ German Research Centre for Geosciences,
  Centre for GeoInformation Technology
  14473 Potsdam, Germany,
  Phone: +49- 331-288-1702
  E-Mail: jens.klump@gfz-potsdam.de

- Olaf Kolditz
  Helmholtz- German Research Centre for Environmental Research – UFZ
  Department Environmental Informatics
  Permoserstraße 15
  04318 Leipzig
  Phone: +49-341-235-1250
  E-Mail: olaf.kolditz@ufz.de

- Frank Neidl,
  Karlsruhe Institute of Technology (KIT),
  Institute of Meteorology and Climate Research – Atmospheric Environmental Research (IMK-IFU),
  82467 Garmisch-Partenkirchen, Germany,
  Phone: +49-8821-183-251
  E-Mail: frank.neidl@kit.edu

- Karsten Rink
  Helmholtz-Centre for Environmental Research – UFZ
  Department Environmental Informatics
  Permoserstraße 15
  04318 Leipzig, Germany
  Phone: +49-341-235-1067
  E-Mail: karsten.rink@ufz.de

- Thomas Schnicke
  Helmholtz Centre for Environmental Research – UFZ
  Scientific and Commercial Data Processing / Scientific Computing and Data Management Group
  Permoserstr. 15,
  04318 Leipzig, Germany
  Phone: +49-341-2351958
  E-Mail: Thomas.Schnicke@ufz.de

- Matthias Schroeder
  GFZ German Research Centre for Geosciences,
  Centre for GeoInformation Technology
  14473 Potsdam, Germany,
  Phone: +49- 331-288-1694
  E-Mail: matthias.schroeder@gfz-potsdam.de

- Jürgen Sorg
  Research Centre Juelich
  Institute for Bio and Geosciences – Agrosphere (IBG-3)
  52425 Jülich, Germany,
  Phone +49-2461-61-5535
  E-Mail: j.sorg@fz-juelich.de

- Vivien Stender
  GFZ German Research Centre for Geosciences,
  Centre for GeoInformation Technology
  14473 Potsdam, Germany,
  Phone: +49- 331-288-28717
  E-Mail: vivien.stender@gfz-potsdam.de

- Ute Wollschläger
  Helmholtz Centre for Environmental Research – UFZ
  Department Monitoring & Exploration Technologies
  Permoserstraße 15
  04318 Leipzig
  Phone: +49-341-235-1995
  E-Mail: ute.wollschlaeger@ufz.de

A 1.2 Institutional Data Managers

DLR: N.N.

FZJ: Ralf Kunkel
52425 Jülich, Germany
Phone: ++49-2461-613262
E-Mail: r.kunkel@fz-juelich.de

GFZ: Vivien Stender
14473 Potsdam, Germany,
Phone: +49- 331-288-28717
E-Mail: vivien.stender@gfz-potsdam.de

HMGU: N.N.

IMK-IFU: N.N.

UFZ: N.N.