D16.4 Report on Use Cases, Requirements, Metadata and Interoperability of WP 16

Document information Summary

Date	03-29-2018
Document title:	D16.4 Report on Use Cases, Requirements, Metadata and Interoperability of WP 16
Leader Partner	UU-team@WP16
Main Author(s):	O. Lange
Contributing author(s):	M. Rosenau (GFZ), M. Dekkers (UU), L. Sagnotti (INGV), F. Funiciello (Roma Tre)
Reviewer(s):	E. Calignano
Approved by:	UU-team@WP16
Target audiences:	Technical staff at EPOS WP's 6 and 7
Keywords:	DDSS, use cases, interoperability, metadata
Deliverable nature:	Upload to intranet
Dissemination level:	Final
Delivery date:	03-29-2018
Version:	3.0

TABLE OF CONTENTS

1 SUMMARY

This document describes the EPOS Thematic Core Service *Multi-scale laboratories* (EPOS Work Package 16), a community built around a wide range of world-class laboratory infrastructures. The community was set up from scratch and efforts were made to implement an environment that aimed at providing the means to engage for every lab that could provide highly relevant scientific data, whatever the scale of the lab and local available resources and infrastructure. Therefore, the overall solution with IT components and data policies is balanced upon this goal and was designed in close collaboration with the field.

The first section of this document discusses the chosen strategy, the role of metadata standards in relation to data publications, and the naturally resulting information architecture for the TCS. It is clarified how data publications form the core entities when sharing relevant scientific data. Consequently, the DDSS definitions are setup through the categorization of these publications into relevant scientific sub domains. In the second section the priority list for M24 is presented, thereby pointing at its dependency on components from the architecture.

The second half of this document is dedicated to the description of use cases related to some already accessible data services. The discussed cases apply to the discovery of published datasets that are uniquely identified by DOIs.

1. Introduction

WP16 (Multi-scale laboratories) is a community that includes a wide range of world-class laboratory infrastructures and that provides a cross-disciplinary, though coherent platform for virtual access to data and physical access to labs. The length scales encompassed by the infrastructures include ranges from the nano- and micrometre levels (electron microscopy and micro-beam analysis) to the scale of experiments on centimetre sized samples, and to analogue model experiments simulating the reservoir scale, the basin scale, and the plate scale. Currently, many of the produced data remain inaccessible and/or poorly preserved. However, the data produced at the participating laboratories are crucial to serving society's need for geo-resources exploration and for protection against geo-hazards. To model resource formation and system behaviour during exploitation, we need an understanding from the molecular to the continental scale, based on experimental and analytical data.

The TCS disseminates these highly relevant laboratory data in the form of sustainable data publications, thereby solving the problems of poor preservation and accessibility. The next sections are concerned with the relevant aspects of setting up an infrastructure for this dissemination: the classification of data into data products, the development of a community metadata standard, and the design of an architecture which connects transparently with the EPOS-ICS. At the end of this chapter contact details can be found in case of further questions about the discussed architecture.

1.1 Cataloguing DDSS's in WP 16.

The Multi-scale Laboratories TCS provides data products in the form of collections of published experimental research data. The resulting DDSSs concern level 0-2 data products that are all treated as constituted of *data publications*, i.e. datasets that are identified with DOIs and described and citable through standardized metadata. Incidentally, it may appear that a specific DDSS concerns rather a database instead of a downloadable individual dataset, still the description and access are managed through the descriptive and contextual metadata and the DOI.

This characterization of the TCS DDSSs as data publications has two profound implications:

1. When every data publication is considered as a separate DDSS, then the number of DDSSs will be too large and continue to increase.

2. When the service that provides access to all data publications is considered as the only DDSS that provides access to data, then it will become difficult to divide the implementation phase into manageable stages. Furthermore, in that case the DDSS list will not reflect the richness of the TCS.

Therefore, the following was decided regarding the high-level requirements for the TCS:

RM1. A fixed DDSS list must contain a scientifically relevant division into subdomains.

RM2. The TCS must provide components (services) supporting data publications, i.e. repositories and a catalogue. These services must have a clear connection to the separate subdomains.

In the EPOS implementation proposal, several priority topics for laboratory data were already chosen, including *analytical and properties data*, *experimental rock physics data*, *analogue modelling data*, and *paleomagnetic data*.

As a first step in further cataloguing the available data (and possibly accompanying available infrastructures for dissemination), several investigations were carried out:

I. a partner contribution and capacity survey

II. a survey on trans-national-access arrangements (not directly related to the DDSS data classification)

III. a survey on the different scientific topics covered by the lab community, plus the type classification of the data produced

IV. an investigation of metadata standards and licensing mechanisms in use

The results of these surveys provided input for the definition of the DDSS priority list for implementation. At an internal WP16 meeting in February 2017 in Utrecht, in accordance with requirement RM1, a first scientific classification into subdomains and DDSSs was agreed on by the TCS community. The following table shows the DDSS definitions that *apply to data provision and that are linked to a single subdomain*, together with their priority statuses that are registered in the EPOS DDSS Master Table:

Subdomain	**DDSS name** **(type ∊ {0,1,2})**	**#DDSS**	**Priority M24**
Geochemical data	Geochemical data (elemental and isotope geochemistry)	WP16-DDSS-001	medium
Rock and melt physical properties	Rock and melt physical properties data	WP16-DDSS-002	high
Analogue models on geologic processes	Property data of analogue modelling materials	WP16-DDSS-003	high
Analogue models on geologic processes	Analogue modelling results	WP16-DDSS-004	high
Analogue models on geologic processes	Software tools	WP16-DDSS-005	low
Paleomagnetic and magnetic data	Paleomagnetic data	WP16-DDSS-006	medium
Paleomagnetic and magnetic data	Magnetic susceptibility data	WP16-DDSS-007	medium
Paleomagnetic and magnetic data	Software tools	WP16-DDSS-008	low

Table 1: Subdomains and DDSSs

To provide access to these data, necessary additional services are declared as DDSS:

DDSS name	Description	#DDSS	Priority M24
TCS catalog	Provides 4 separate webservices for the 4 different subdomains. These web services connect directly to the ICS-C for the exchange of EPOS-DCAT-AP metadata. The catalog entails an end user portal for discovery and description of datasets from the DDSSs 001-008. Metadata is the catalog is a superset of the EPOS-DCAT-AP metadata exchanged with the ICS-C.	WP16-DDSS-012	high
GFZ data repository	Sustainable storage of and access to datasets from DDSSs 001-008. Landing pages are resolved by DOIs. The environment provides a user-friendly interface for dataset description by a scientist (metadata editor).	WP16-DDSS-013	high
UU data repository	Sustainable storage of and access to datasets from DDSSs 001-008. Landing pages are resolved by DOIs. The environment provides a user-friendly interface for dataset description by a scientist (metadata editor).	WP16-DDSS-014	high

Table 2: service oriented DDSSs

These 9 DDSSs with priorities high/medium are logically connected in the following architectural picture:

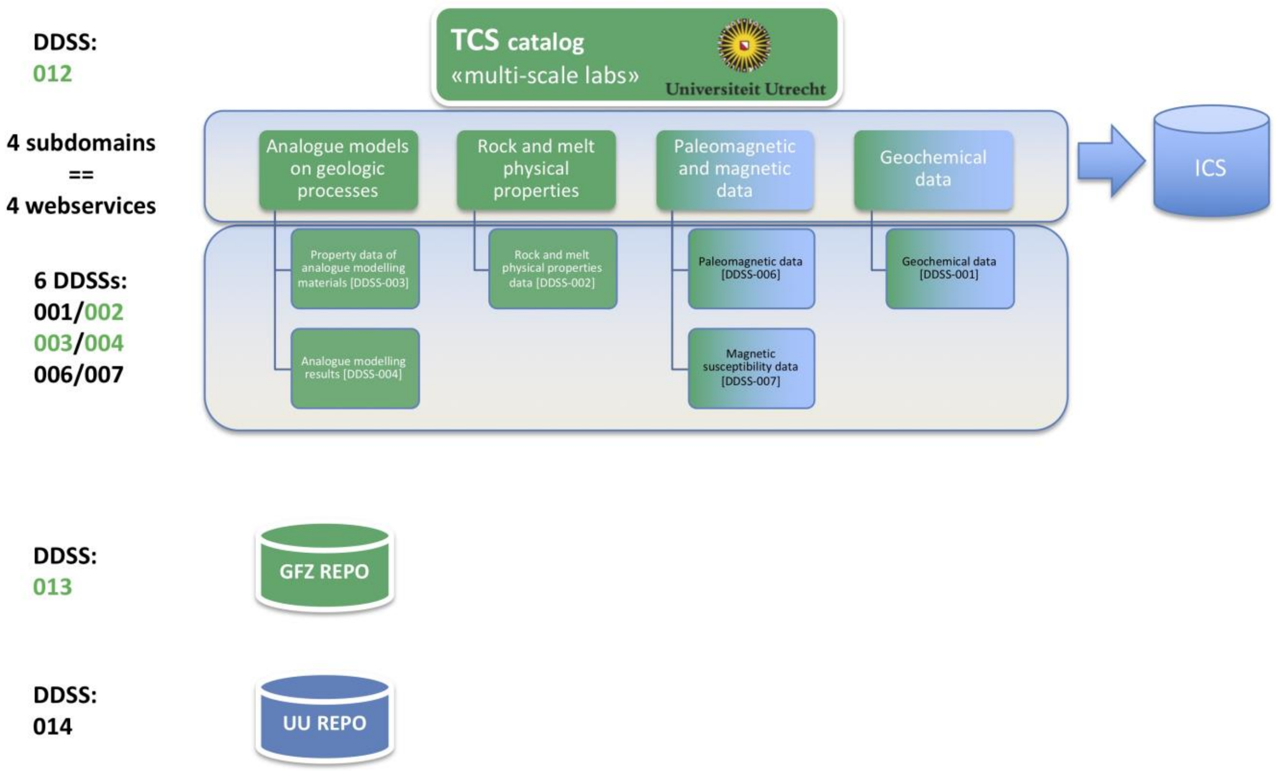
DDSS:
012
TCS catalog
«multi-scale labs»
Universiteit Utrecht
4 subdomains
==
4 webservices
Analogue models on geologic processes
Rock and melt physical properties
Paleomagnetic and magnetic data
Geochemical data
ICS
6 DDSSs:
001/002
003/004
006/007
Property data of analogue modelling materials [DDSS-003]
Rock and melt physical properties data [DDSS-002]
Paleomagnetic data [DDSS-006]
Geochemical data [DDSS-001]
Analogue modelling results [DDSS-004]
Magnetic susceptibility data [DDSS-007]
DDSS:
013
GFZ REPO
DDSS:
014
UU REPO

Figure 1: connections between DDSSs and services ('green' applies to priority 'high')

1.2 Using standards and best practices: development of the *metadata* scheme

As a starting community WP16 had the great opportunity to design the TCS in a fully logical and compatible way, both from the metadata and IT perspectives. Therefore, the crucial first step was the development of a homogenous cross-disciplinary metadata scheme for discovery and contextual description of data emerging from the multi-scale labs.

WP 16 brings together representatives of experimental communities having a hugely different set of instrumentations, scientific background and expertise. Even though all the laboratories work within the broad field of geomaterials- and geological processes characterizations, a coherent and effective structure for data inventory was lacking. On the other hand, because the metadata descriptions for all the community provided data were eventually to be used at the level of the central EPOS-ICS catalogue, both discipline agnostic information and specialized contextual descriptions had to be provided in a standardized way. So, interdisciplinary usage and findability implied a necessary usage of proven metadata standards and vocabularies wherever possible. Because of this, DataCite 3.2 (domain agnostic standard for descriptive metadata) and INSPIRE/ISO19115 were taken together as the logical starting point for describing the community data.

As explained above, the data provided through the TCS are at best characterized as *data publications*, i.e. citable data underlying experimental research, and uniquely identifiable as data sets published with a DOI. Following the FAIR principles for Open Science (data must be *F*indable, *A*ccessible, *I*nteroperable, and *R*eusable), and the acknowledgment that most of the contextual metadata had to be supplied by researchers themselves, the following high-level requirements naturally emerged concerning the metadata scheme:

RM3. Community standards for both descriptive and contextual metadata must be followed wherever possible.

RM4. Providing metadata must be possible in a user-friendly manner.

RM5. The metadata fields and the quality of its values must be sufficient for a first assessment of the usefulness of a data set at the discovery level.

RM6. Accessibility and terms of use for data must be unambiguous through the use of administrative metadata.

Following these requirements, the following decisions were made:

1. INSPIRE/ISO19115 must be taken as the standard for the contextual TCS metadata.

2. To support external references (structural metadata for publications, institutions, etc.) DataCite 3.2 must be included.

3. Data citations must be supported with the use of DOI's and the implied mandatory descriptive fields from DataCite 3.2.

4. Where the common controlled vocabularies in use for the Earth sciences (e.g. NASA GCMD, INSPIRE, ANDS) appear not to fulfil the needs for contextual information, extensions must be provided using additional *keywords*. These keywords must be provided and maintained as ***controlled vocabularies*** within its own *namespace*.

5. These extended keywords must be provided by researchers using a dedicated user-friendly metadata editor. Further descriptive and administrative metadata are supposed to be generated automatically wherever applicable.

With this strategy, WP 16 was able to design a flexible scheme of metadata that could be associated with raw data products (level 0-2), and which provided a coherent structure for all the lab-related disciplines. The scheme was designed to provide basic, but freely implementable, information on the data products to enhance data discoverability and classification. Since researchers themselves primarily do the classification of lab data, the proposed scheme had been designed to be as less time consuming as possible: the data product is assigned with a DOI, which is tagged by a set of standardized metadata, mostly in the form of keywords, which in turn are the target of the query from the Central EPOS Hub.

So, the classification form that was developed in collaboration with lab researchers subdivides the metadata in three broad groups:

1) Metadata automatically associated with the data product during the researcher's login;

2) Generic metadata, common to all the disciplines;

3) Metadata related to the specific research field, designed to have keywords nested in hierarchic levels to allow both the desired level of complexity in the description of the data products and a time-effective classification by the researcher.

The metadata scheme thus designed is expandable in complexity, for example with the possibility to extend the vocabularies. Furthermore, we have added the possibility to copy-paste metadata from one product to another to speed up the classification process for the researcher.

Eventually, the TCS metadata is exposed in a *fully standardized* DCAT-based manner. This was confirmed and tested at the WP6/7 hackathon meeting, which preceded the EPOS second integration workshop in Prague in February/March 2017.

1.3 Standards and flexibility: overview of a sustainable ICT architecture

Setting up an infrastructure for sustainable data publications concerns the implementation of standard components and the adoption of best practices in use around the world. Central functional ingredients are especially the *sustainable storage of data*, *assignment of DOI's*, *editing and exposition of metadata*, and *maintenance and editorial review*. Besides these standard functionalities, for the TCS the presumed infrastructure must fulfil a very important additional high-level requirement:

RM7. Data publication with TCS compliant metadata must be possible for every engaged lab, independent of locally available resources (IT infrastructure, information specialists).

This requirement forms the basis for the rationale behind the chosen infrastructure. After all, the Multi-scale laboratories TCS concerns the engagement of research labs into a growing community. Populating the infrastructure with data, and the community with participants, together form the main goal for the TCS. The following architecture matches these goals and foresees in flexible scenarios for growth and ambition:

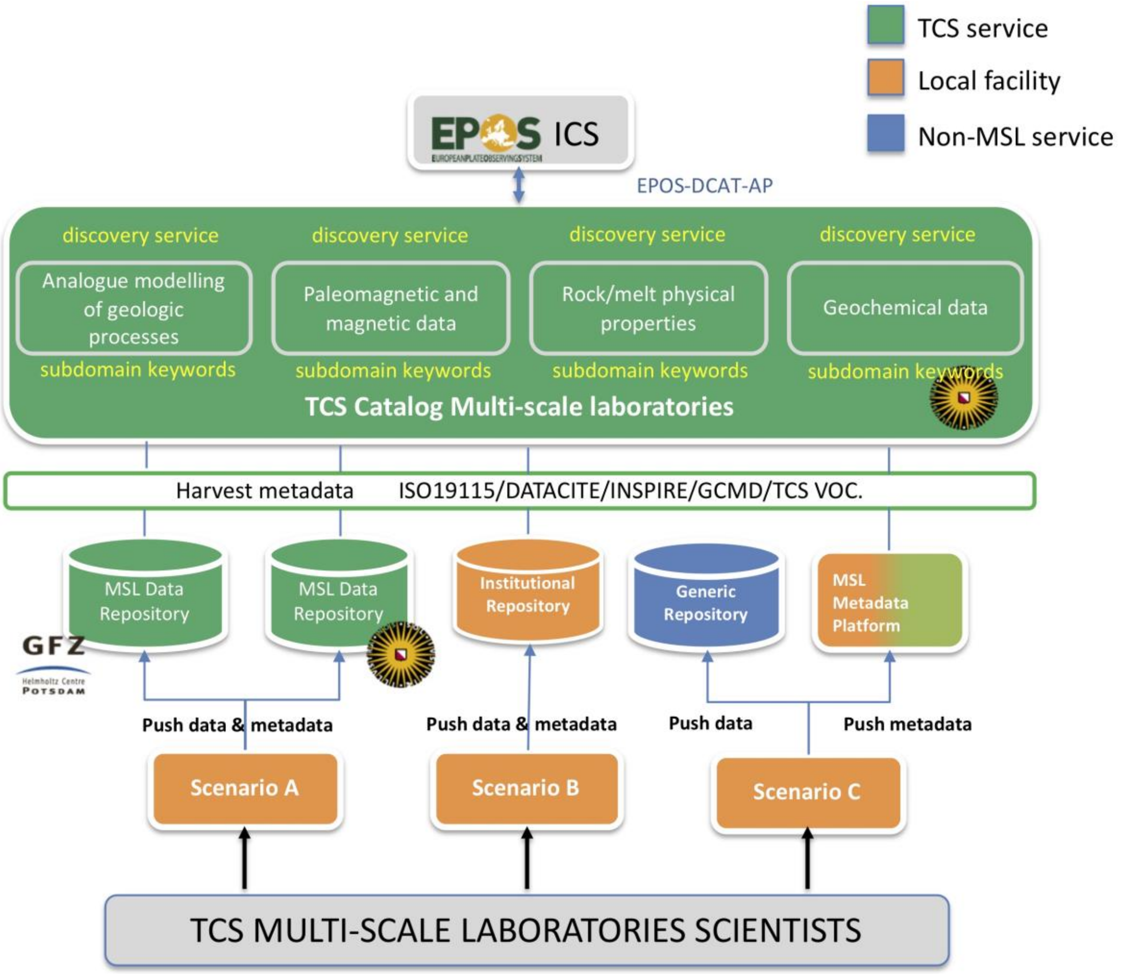

Figure 2: Multi Scale Laboratories Data infrastructure

Although the illustration is self-explaining to a certain level, some aspects are worth emphasizing:

- The infrastructure provides 3 scenarios for laboratories to engage as a data provider:

    1. through the usage of TCS-wide repositories with facilities for assigning DOI's and adding TCS compliant metadata,

    2. through the usage of a local repository that can handle TCS compliant metadata,

    3. through the usage of a general repository for data storage, in combination with the TCS metadata editor for the generation of compliant metadata. Engaging labs have the possibility to grow from a localized facility towards a TCS-service.

- Local research infrastructures are only indirectly connected to the ICS; they share metadata with ICS-C via the TCS catalogue.

- The DOI's which are shared with the ICS catalogue, refer to the landing pages of the data sets (or software tools/databases).

- The DDSSs 001-008 that were described in the former section, are handled in the same way, and differ only by the values of their metadata. As a result, *testing a DDSS concerns testing the overall infrastructure, including the applicable part of the metadata scheme (i.e. the keywords extension).*

- The (virtual) TCS metadata platform (MMP) contains metadata which is preferably stored with the data. This is not always possible (fi. in case of a national repository), and in these cases the metadata will be stored close to the TCS catalogue.

- In this overview, the physical accessible TCS services concern the repositories at Utrecht University and GFZ Potsdam, and the CKAN catalogue at Utrecht University.

- The individual repository implementations can differ in their underlying techniques (for example, the Utrecht repository relies on iRODS technology, which is not the case at Potsdam). They all provide landing pages for the locally published datasets.

- The TCS CKAN catalogue externally exposes metadata in the EPOS-DCAT-AP format as a subset of a larger collection of metadata in ISO-format. The externally exposed metadata was confirmed at the workshop in Prague to be successfully transformable to the necessary JSON-format for further mapping into CERIF.

1.4 Summary

In the previous sections both the metadata scheme and the infrastructure of the TCS were discussed through their relations to some high-level criteria, that we sum up here again:

RM1. The DDSS list is defined upon a scientifically relevant division into subdomains and necessary servcies.

RM2. The TCS will provide the necessary services for the full support of data publications.

RM3. Community standards for both descriptive and contextual metadata will be followed wherever possible.

RM4. Providing metadata by researchers must be possible in a user-friendly manner.

RM5. The metadata fields and the quality of its values must be sufficient for a first assessment of the usefulness of a data set at the discovery level.

RM6. Accessibility and terms of use for data must be unambiguous by using administrative metadata.

RM7. Data publications through the TCS must be possible for every lab, independent of the availability of local resources.

The infrastructure, which is in line with these constraints, was sketched in the previous section, and it contains the following fixed components:

TCS Service/infrastructure component	**Hosted at**	**Operational at M24**
2 repository services at the TCS level (number can be extended) (Both include metadata editor for editing TCS compliant metadata)	GFZ Potsdam Utrecht University	X (M24-M30)
DOI assignment + minting DOI assignment	GFZ Potsdam Utrecht University	X (M24-M32)
TCS catalogue for harvesting and ICS-connection	Utrecht University	X

Table 3 – Components for the support of TCS data publications

1.5 Information and contact

For any information about the TCS infrastructure, please contact

Otto Lange (Utrecht University) – WP16 package leader and IT contact

e-mail: o.a.lange@uu.nl

direct phone: +(31) 6 51 31 7777

2 Priority List of DDSS

As explained thoroughly in the last chapter, the DDSS specifications rely on a subdomain dependent distribution of data publications, i.e. data sets with DOI's and standardized metadata, and which are all handled in a uniform functional manner through the TCS infrastructure. So, prioritization and implementation of the TCS DDSSs automatically implies the implementation of the necessary parts of the architecture that was discussed. Furthermore, validating and testing the different DDSSs concerns populating one of the repositories with the applicable data, and assessing the keyword extensions (TCS vocabularies) of the metadata scheme. The following table contains the DDSSs which were available by M24, i.e. at the time of the Sept 2017 Technical Readiness Assessment (TRA):

Subdomain	**DDSS (data publication category)**	**Priority: available at M24**
Rock and melt physical properties	Rock and melt physical properties [002] Data and metadata ready	Yes
Analogue models on geologic processes	Property data of analogue modelling materials [003] Data and metadata ready	Yes
Analogue models on geologic processes	Analogue modelling results [004] Data and metadata ready	Yes
All subdomains	TCS catalog [012] Succesfully passed the TRA, i.e. connected as a service with the ICS-C	Yes
All subdomains	GFZ repository [013]	Yes

Table 4 – DDSS priority list

It must be stressed that from the high priority DDSSs ***only DDSS-WP16-012 (TCS catalog) could be assessed during the TRA***. The reason is that TRA concerned a test of web services integration into the ICS-C portal. However, the TCS catalog DDSS-012 is the only DDSS that directly connects as a service to the ICS-C. This is an immediate result of the overall architecture (see figures 1 and 2). During the TRA is became apparent that the EPOS-DCAT-AP schema in use expected a cardinality of '1' for the value 'subdomain' when describing a web service. This implies that every web service serves a single subdomain. However, the TCS portal provided at the time a single parameterised web service, through which values for subdomains could be provided as parameters. Therefore, it became necessary to split the single API into 4 separate subdomain dependent web services, i.e. the current situation as sketched in figure 1. It is agreed with WP6/7 both at the Bucharest meeting and again at the Lisbon workshop that the connection in the form of 4 separate services will again be tested in the announced WP7 collaboration time-window of April-May 2018.

Obviously, the TCS catalogue has high priority because it is the necessary component for combining individual data publications into the DDSS groups that are defined in table 1. Therefore, the subsequent sections will provide information about the following:

1. The TCS catalogue and its role concerning metadata exchange with both the ICS-catalogue and the TCS repositories.
2. Some example data publications in the form of published datasets from the subdomain '*Property data of analogue modelling materials*' (see Table 4), which will illustrate metadata standards in use, and the format of landing pages that are resolved via DOI's.

2.1 The TCS catalogue

According to the architecture sketched in both figures 1 and 2, the TCS catalogue is the 'single point of access' for harvesting TCS community metadata into EPOS-ICS (CERIF). I.e. all descriptions of data in the EPOS-ICS are retrieved via EPOS-DCAT-AP based metadata synchronization with the TCS catalogue.

The catalogue is 'passive' in its role, i.e. metadata is not published into the CKAN database, but the latter is rather filled through gathering metadata via OAI-PMH endpoints at the underlying repositories. The choice for CKAN is based on best practices within Utrecht University and EUDAT (B2Find), and its ability to expose standard metadata in DCAT-AP format (amongst others).

2.2 DDSSs: the structure of data publications

The following table contains some examples of individual data publications from the WP16-DDSS-003, *Property data of analogue modelling materials* (subdomain *Analogue models on geologic processes*). All the data examples in table 4 are comprised of collected data publications, and their descriptions are made accessible in a uniform way via the standardized information that is exposed via the TCS catalogue.

DDSS type: data publications concerning

PROPERTY DATA OF ANALOGUE MODELLING MATERIALS

Available at

- *GFZ Data Services*

Type of data:

Measured and processed Datasets (Data Products): Rock analogue material properties

Type of resource: friction data

Authors: Klinkmüller, M.; Schreurs, G.; Rosenau, M.

Name: GeoMod2008 materials benchmark: The ring shear test dataset

URL: http://doi.org/10.5880/GFZ.4.1.2016.002

Type of resource: rheology data

Authors: Rudolf, M.; Boutelier, D.; Rosenau, M.; Schreurs, G.; Oncken, O.

Name: Supplement to: Rheological benchmark of silicone oils used for analog modeling of short- and long-term lithospheric deformation.

URL: http://doi.org/10.5880/GFZ.4.1.2016.001

Type of resource: triax data

Authors: Klinkmüller, M.; Schreurs, G.; Rosenau, M.

Name: GeoMod2008 materials benchmark: The axial test dataset.

URL: http://doi.org/10.5880/GFZ.4.1.2016.006

Type of resource: SEM imagery

Authors: Klinkmüller, M.; Kemnitz, H.; Schreurs, G.; Rosenau, M.

Name: GeoMod2008 materials benchmark: The SEM image dataset.

URL: http://doi.org/10.5880/GFZ.4.1.2016.004

Type of resource: sieve data

Authors: Klinkmüller, M.; Schreurs, G.; Rosenau, M.

Name: GeoMod2008 materials benchmark: The sieve dataset.

URL: http://doi.org/10.5880/GFZ.4.1.2016.003

Format(s) of the data (if applicable)

General format of the datasets:

Downloadable compressed (zip) data packages, consisting of

a) data files

b) overview and explanations (pdf).

c) file listing in xslx (for the suplementary rheology data)

<u>Data formats</u>:

- Binary: tdms, lfx (raw time series data) – in future
- Ascii: txt (raw time series and converted/processed/analyzed data)
- xls (converted/processed/analyzed data)
- pdf (visualized data)
- Image files: tif (raw image data)

Metadata standards used

- ISO19115/ISO19139 in INSPIRE profile using free keywords (http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020)
- NASA GCMD Science Keywords (http:// gcmd.nasa.gov/learn/keyword_list.html)
- GEMET Thesaurus
- Datacite Metadata Schema (v3.1) providing mime-types of the files and details about the authors and their institions/affiliations (http://schema.datacite.org/meta/kernel-3/index.html) (schema: http://schema.datacite.org/meta/kernel-3/metadata.xsd)
- dif version 9.8.2 (https://earthdata.nasa.gov/standards/directory-interchange-format-dif-standard)
- Dublin Core (DC) (see OAI-PMH entry in the APIs section)

The data descriptions based on the above standards are available as XML via the landing pages of the data products, which in return are accessible through the DOI's given above.

APIs used to provide discovery and access to the DDSS

- Datasets are accessible through *doi-landing pages*. The datasets must be downloaded via this web page. There is no direct pointer to the files themselves.
- Metadata about the datasets are accessible through an OpenSearch interface and an OAI-PMH endpoint in the metadata profiles mentioned above:.
- For humans, search is available here: (*human readable*)
    - http://doidb.wdc-terra.org/search/public/ui?&sort=updated+desc&q=subject%3Aepos+and%20subject%3Aanalog*
- For machines, providing related material and other important information: (*machine readable*)

- XML: http://doidb.wdc-terra.org/search/public/api?q=subject%3Aepos+and%20subject%3Aanalog*&fl=relatedIdentifier,doi,creator,title,publisher,publicationYear,datacentre,subject,description,contributor,date,size,format,rights,relatedIdentfier&fq=is_active:true&fq=has_metadata:true&rows=10&sort=updated+desc&wt=xml&indent=true

- CSV: http://doidb.wdc-terra.org/search/public/api?q=subject%3Aepos+and%20subject%3Aanalog*&fl=relatedIdentifier,doi,creator,title,publisher,publicationYear,datacentre,subject,description,contributor,date,size,format,rights,relatedIdentfier&fq=is_active:true&fq=has_metadata:true&rows=10&sort=updated+desc&wt=csv&indent=true

- JSON: http://doidb.wdc-terra.org/search/public/api?q=subject%3Aepos+and%20subject%3Aanalog*&fl=relatedIdentifier,doi,creator,title,publisher,publicationYear,datacentre,subject,description,contributor,date,size,format,rights,relatedIdentfier&fq=is_active:true&fq=has_metadata:true&rows=10&sort=updated+desc&wt=json&indent=true

- OAI-PMH: http://doidb.wdc-terra.org/oaip/oai?verb=ListRecords&metadataPrefix=oai_datacite&set=~P3E9c3ViamVjdCUzQWVwb3MrYW5kK3N1YmplY3QlM0FhbmFsb2cq

Authentication, Authorization, Accounting Infrastructure (AAAI)

- Data are openly available, i.e. open access without embargo applied and under CC-BY open license.

- There is no AAAI in place and there are no plans to implement such a system.

Data policy

Licensing:

These datasets are licensed under *CC BY 4.0*

(Human readable: https://creativecommons.org/licenses/by/4.0/)

(Formal legal code: https://creativecommons.org/licenses/by/4.0/legalcode)

The datasets are published as supplements to peer reviewed journal articles, available after review, and without embargo.

Other technical details

In the case where the following topics are applicable to this DDSS only (and are not TCS-wide) may you provide technical details and roadmap for implementation?

1. Data curation system (specify which system. Is it interoperable?)

Repository of GFZ Data Services (http://doi.org/10.3390/ijgi5030025) / Yes, it is interoperable

2. Data provenance (do you track provenance? Which standards are you using? Which software?)

Versioning, editorial review, and source references.

3. Identification (do you have a system to assign identifiers to this DDSS? Which is your roadmap to this respect? Please provide technical details)

DOI's are assigned o *all data publications.*

4. Do you provide processing features for this DDSS? Which ones? Which standard are you using to provide programmatic access to your processing services by ICS?

Not yet. In case we provide binary files, a data viewer will be available along with the data.

Roadmap for implementation

Please provide a roadmap for the implementation of this DDSS. Alternatively, you may provide a separate diagram covering all DDSS (e.g. gantt).

This DDSS is implemented. (see table 4)

Table 5: examples of data publications from DDSS-003

3 TCS roadmap

As was already explained in the introduction, the mission of the TCS is to build a community of European high-quality labs. This implies on the one hand a necessary infrastructure through which every lab can engage, and on the other hand activities to populate this infrastructure with scientifically relevant data. The challenge is therefore two-fold: first, the infrastructure must support the full process of data publication and dissemination via the EPOS-ICS catalogue, and secondly, the community must be activated in supplying data. Therefore, the following roadmap is followed in realizing the TCS goals:

M1-M24: necessary components of basic infrastructure for data publications have become operational

- Repository operational at GFZ
- DOI services operational at GFZ
- Metadata editor operational at GFZ
- TCS catalogue operational at Utrecht University
- Metadata scheme finalized for two subdomains (*Rock and melt physical properties* and *Analogue models on geologic processes*)
- Populated with datasets

M24-M32: extending infrastructure with extra components (extension of service providers)

- Repository operational at Utrecht University[1]
- DOI services operational at Utrecht University
- April-May 2018 TCS catalog web service description refined for additional use cases at ICS-C

M24-M32: further populating infrastructure with data.

(At the internal WP16 meetings in February 2017 in Utrecht and October 2017 in Rome, participating labs have been instructed in workshops on how to start the publication of data. These instructions concern an ongoing effort.)

[1] The repository at Utrecht University is part of a large institutional effort for setting up a sustainable environment for data curation and publication. The system is built on iRODS technology and supports, amongst others, secure mechanisms for versioning, collaboration, WebDAV access, etc. It follows the architectural paradigm of the Open Archival Information System. Currently the environment is used under the name YoDa ("Your Data") within large interdisciplinary projects for the management of highly sensitive data.

4 Data Management Plan (DMP)

The TCS data management plan is in development. Finalization is planned for June 2018.

5 Use cases

Several conceptual use cases for interdisciplinary usage of the TCS datasets are imaginable: data on volcanic ash could be utilized by the aviation industry, meteorological institutes, and governments in decision making on the response to volcanic ash eruptions. The experimental rock physics and analogue model data could be used by scientists modelling sedimentary basin formation, and in the exploration for unconventional resources and geothermal energy. Paleomagnetic data could be used in charting geo-hazard frequency.

Some more specific and detailed technical use cases are given below. These descriptions apply to datasets that are already accessible via their applicable DOIs.

5.1 Use case 1: material parameters

Use case name/topic:

Retrieve material parameters for setting up analog experiments

Use case domain

This use case is focused on the domain of *Geodynamics*

Use case description.

Problem:

As an <analog modeler> Graham wants to < decide>< which material (brittle, ductile) is suitable for his experimental setup>.

Elucidation:

Friction and viscosity are crucial experimental parameters that can be influenced by a right choice of used materials.

Persona:

Graham is a geologist. He usually uses a centrifuge for analogue modelling, which requires rather stiff materials (clay, plasticine). As his centrifuge broke down he now needs weaker materials to setup an experiment under normal gravity conditions.

Supposed context:

Graham knows that EPOS might provide useful information for his setup. He is however not certain about the form of the available information (i.e. of the data product), so he does not know whether

- *he will learn about the availability of data sets that can be used by him for further analysis to decide which materials are suitable, or*
- *that he can consult a database or other piece of software to directly retrieve raw material values.*

So, he is unaware of the representation of the data that will be valuable to him.

<u>Result</u>:

Graham has assessed and downloaded the data sets (and associated papers). He has decided on a suitable material, and moreover he has gained knowledge about the materials which he didn't have before.

Actors involved in the use case A list of the actors who communicate with this use case.

- *Researcher (system user)*
    - *Analog modeler*

Priority: *Medium*

Pre-conditions:

- *<analog modeler> is logged in as a named user.*

Flow of events – user view

Basic sequences and needed steps (user view) – *<u>Full text search without pre-filtering</u>*

1. *<analog modeler> is going to search for material properties. He searches directly for "viscosity material properties" with the aid of a generic text search box.*
2. *<analog modeler> applies a filter to the rather large result set by choosing from a keywords facet the combination of string values "viscoelasticity" and "analog models".*
3. *<analog modeler> clicks in the updated result set on the entrance "GeoMod2008 materials benchmark: The ring shear test dataset", which brings him to the landing page of this data product.*
4. *<analog modeler> assesses this data set by a) reading the abstract, b) viewing the inline iso19115 metadata, and c) reading the explanation of the dataset which is provided as a downloadable pdf ("Explanations for the RST dataset.pdf").*
5. *Based on his analysis and the open availability of the dataset <analog modeler> downloads the compressed data package ("RST-data.zip").*

Alternatives for steps 1 and 2 (user view) – *<u>Full text search with filter on type of requested data</u>*

1. *<analog modeler> is going to search for material properties. Therefore, he applies a search filter "material properties" (chosen from a topic list) as a constraint on a search query that he is going to execute.*
2. *<analog modeler> now searches free text with the search string "friction viscosity".*

Alternative for steps 1 and 2 (user view) – *<u>Searching through filtering without full text search</u>*

1.b *<analog modeler> is going to search for material properties. Therefore, he applies a search* *filter "material properties" (chosen from a topic list) that directly leads him to a result set.*

2.b *<analog modeler> refines the result set by applying the filters "viscosity" and "friction" (within the facet "property type"), and "analog models" (from a list of keywords), again directly leading to new results sets.*

System workflow - system view

Full text search without pre-filtering

1. *The GUI receives the input: new full text search*
   a. *The system connects to the full text index and executes the query "viscosity material properties" over all indexed fields*
      i. *The system makes use of fuzzy search techniques and/or vocabularies to relate "viscosity" to "viscoelasticity".*
   b. *The system returns a results page*
      i. *That is shown in the GUI*
      ii. *That contains filters (facets) that apply to the underlying result set (i.e. all results have a keyword field, and "keyword" is contained as a restricted list of values upon which further refinements can be made; values that do not apply to some individual result are not in the refinement list).*
2. *The GUI receives the input: applied filter*
   a. *The system connects to the full text index and executes the query <keywords: "viscoelasticity" AND keywords: "analog models"> onto the last result set.*
   b. *The system returns a result set that is a refinement of the former one.*
3. *The GUI receives the input: hyperlink clicked*
   a. *The URL is followed (possibly in a new target tab of the browser)*
   b. *User session remains active*
4. *Steps 4 and 5 are not controlled by the system. However, the user may return to update his choice of filter values, in which case action 2 is executed again. (A request for a new search may or may not reset applied filtering.)*

Full text search with filter on type of requested data

1.a *The GUI receives the input: filter applied*

a. *Search filter "material properties" receives value "material properties".*

2.a *The GUI receives the input: new full text search with filter applied*

a. *The system connects to the full text index and executes the query < ("friction" OR "viscosity") AND (topic: "material properties")>*

i. *The system makes use of fuzzy search techniques and/or controlled vocabularies to relate "viscosity" to "viscoelasticity".*

ii. *Search is over all indexed fields, except "material properties", which is only sought within the field "topic".*

b. *The system returns a results page*

i. *That is shown in the GUI*

3.a The GUI receives the input: hyperlink clicked

a. The URL is followed (possibly in a new target tab of the browser)

b. User session remains active

Steps 4 and 5 are not controlled by the system. However, the user may return to update his choice of filter values, in which case action 1 is executed again.

<u>Searching through filtering without full text search</u>

1.b *The GUI receives the input: new search with filter applied*

a. *The system connects to the full text index and executes the query <topic: "material properties">*

i. *Search is over the field "topic".*

b. *The system returns a results page*

i. *That is shown in the GUI*

ii. *That contains filters (facets) that apply to the underlying result set (i.e. all results have a keyword field, and "keyword" is contained as a restricted list of values upon which further refinements can be made; values that do not apply to some individual result are not in the refinement list).*

2.b *The GUI receives the input: filter applied*

a. The system connects to the full text index and executes the query <keywords: "viscoelasticity" OR keywords: "analog models"> onto the last result set.

b. The system returns a result set that is a refinement of the former one.

3.b The GUI receives the input: hyperlink clicked

a. The URL is followed (possibly in a new target tab of the browser)

b. User session remains active

Steps 4 and 5 are not controlled by the system. However, the user may return to update his choice of filter values, in which case action 1 or 2 is executed again.

Post-conditions

- *<analog modeler> is logged in as a named user.*
- *<analog modeler> owns a search results object*
- *<analog modeler> owns a 'last query executed'*
- *<analog modeler> is connected to an active search parameterization profile (with or without constraints).*

Extension Points

None

« Used » Use Cases Determine the systems functionality that might be reused and model this using the <<uses>> relationship. If the use case uses other Use Cases, list them here.

…

Other Requirements This can include non-functional requirements related to the Use Case.

…

(to be filled in by WP7) After the interview: create class and sequence diagram for each use case. Class diagram and sequence diagram.

5.2 Use case 2: sand rock comparison

Use case name/topic:

Comparison of sand with rock

Use case domain

This use case is focused on the domain of *Geo-engineering*

Use case description.

Problem:

As an <engineer> Guido wants to < assess><the stability of his underground storage reservoir>. Therefore he wants to <investigate> the hypothesis that < sand is mechanically comparable to rock>, < depending on the scale of observation>.

Persona:

Guido is a geo-engineer. He worries about the stability of his underground storage reservoir and would like to setup small-scale experiments to test some key issues.

Supposed context:

Guido knows EPOS might provide the information he needs. He is unaware of the representation of data that will be valuable to him.

Result:

Guido has verified that, depending on the scale of observation, sand indeed mechanically behaves comparable to rock.

Actors involved in the use case A list of the actors who communicate with this use case.

- *Researcher (system user)*
    - *Geo-engineer*

Priority: *Medium*

Pre-conditions:

- *< Geo-engineer > is logged in as a named user.*

Flow of events – user view

Basic sequences and needed steps (user view) – *Full text search without pre-filtering (The 'Google Search')*

1. *<Geo-engineer> is going to search for context models of material properties. He searches directly for "sand rock mechanics" with the aid of a generic text search box.*

2. *< Geo-engineer > applies a filter to the rather large result set by choosing from a topic facet the string value "analog models".*

3. *<Geo-engineer > clicks in the updated results set on the entrance "GeoMod2008 materials benchmark: The axial test dataset.", which brings him to the landing page of this data product.*

4. *<Geo-engineer > assesses this data set by a) reading the abstract, b) viewing the inline DataCite metadata, and c) reading the explanation of the dataset which is provided as a downloadable pdf ("Explanations for the AT datasets.pdf").*

5. *Based on his analysis and the open availability of the dataset <Geo-engineer> downloads the compressed data package ("AT-data.zip").*

System workflow - system view

Full text search without pre-filtering

1. *The GUI receives the input: new full text search*

    a. *The system connects to the full text index and executes the query "sand rock mechanics" over all indexed fields*

    b. *The system returns a results page*

        i. *That is shown in the GUI*

        ii. *That contains filters (facets) that apply to the underlying result set (i.e. all results have a topic field, and "topic" is contained as a restricted list of values upon which further refinements can be made; values that do not apply to some individual result are not in the refinement list).*

2. *The GUI receives the input: applied filter*

    a. The system connects to the full text index and executes the query <topic: "analog models"> onto the last result set.

    b. The system returns a result set that is a refinement of the former one.

3. *The GUI receives the input: hyperlink clicked*

    a. The URL is followed (possibly in a new target tab of the browser)

    b. User session remains active

4. *Steps 4 and 5 are not controlled by the system. However, the user may return to update his choice of filter values, in which case action 2 is executed again. (A request for a new search may or may not reset applied filtering.)*

Post-conditions

- *< Geo-engineer > is logged in as a named user.*
- *< Geo-engineer > owns a search results object*
- *< Geo-engineer > owns a 'last query executed'*
- *< Geo-engineer > is connected to an active search parameterization profile (with or without*

constraints).
Extension Points *None*
« Used » Use Cases Determine the systems functionality that might be reused and model this using the <<uses>> relationship. If the use case uses other Use Cases, list them here. …
Other Requirements This can include non-functional requirements related to the Use Case. …
(to be filled in by WP7) After the interview: create class and sequence diagram for each use case. Class diagram and sequence diagram.

6 CONCLUSION

All initial necessary components for the support of data publications have successfully become part of an operational infrastructure at M24. Together with a finalized metadata scheme for two out of four subdomains, the TCS is fully prepared for populating the TCS with data. From M30 on, the infrastructure will be extended with additional repository connections. These apply to the internal TCS repository at UU, and to external repositories in the UK and Spain (NERC and CSIC). All activities are aimed at spreading the available TCS services among different institutes. Metadata for the two remaining subdomains *Paleomagnetic and magnetic data* and *Geochemical data* are planned to become finalized in the Spring of 2018.