Establishing the European Geological Surveys Research Area to deliver a Geological Service for Europe

Authors and affiliation:
Dana Čápová (ČGS),

Olga Moravcová (ČGS),

Lucie Kondrová (ČGS),

Pavla Kramolišová (ČGS)

Martin Hansen (GEUS)

E-mail of lead author:
dana.capova@geology.cz

Deliverable 7.1

Working version Metadatabase

Version: 13-12-2019

This report is part of a project that has received funding by the European Union's Horizon 2020 research and innovation programme under grant agreement number 731166.

Deliverable Data		
Deliverable number	D7.1	
Dissemination level	Public	
Deliverable name	Working version Metadatabase	
Work package	WP7, Developments (central)	
Lead WP/Deliverable beneficiary	GeoZS	
Deliverable status		
Submitted (Author(s))	13/12/2019	Dana Čápová
Verified (WP leader)	13/12/2019	Andrej Vihtelič
Approved (Coordinator)	16/12/2019	Jørgen Tulstrup

GENERAL INTRODUCTION

This deliverable briefly presents the working version of the metadatabase as a part of the GeoERA Information Platform infrastructure (EGDI). It was developed as the central access point to metadata describing in the standardized form all identified digital and structured data resources and other selected information delivered by the 14 scientific GeoERA projects. This report describes the application of the EGDI Metadata Catalogue (MIcKA), available on https://egdi.geology.cz/ and on the project portal http://www.europe-geology.eu/metadata/.

TABLE OF CONTENTS

1 INTRODUCTION TO THE EGDI METADATA

1.1 Metadata of the structured data

The EGDI Metadata Catalogue (MIcKA) is the central access point to metadata concerning structured data on geo-energy, groundwater and raw materials themes provided by the geoscientific GeoERA projects. It provides tools for compilation of those metadata in a standardized format. In order to make the data discoverable in the most efficient way, the catalogue is fully compliant with international standards and supports the distributed system of metadata administration. In order to display a metadata record for which an on-line map service is available, the Metadata Catalogue is integrated into the EGDI Portal http://www.europe-geology.eu/. The catalogue enables systematic discovery, viewing and use of available geological data across Europe. The working version of the EGDI metadata catalogue is operational at https://egdi.geology.cz/ and on the project portal http://www.europe-geology.eu/metadata/.

1.2 Metadata of the unstructured data

It has been decided not to store metadata for unstructured documents in MIcKA, so this deliverable only marginally mentions this problem. The technical solutions will be developed by another team as a part of other tasks of WP5 and WP6. The unstructured data will be stored within the EGDI platform and the data provider will be asked to enter some information about the file during the upload process. This information might depend on the type of the document (pdf, image, csv, etc.). This metadata on documents will include information about which project the file belongs to and relevant keywords. It will be possible (but not required) to geotag the document.

It will be possible to store spatial data that refers to documents in the document repository. In this case there must be added metadata to the EGDI metadata catalogue (MIcKA). This could be a shapefile with the location of test sites as polygons where a pdf-report describing them is attached to each polygon.

There will be a data entry form established for the upload of documents to the document repository. This will be part of the administration module where users can add their data / services to the EGDI portal. The metadata added during upload will be stored in the PostgreSQL database used by the EGDI platform and the documents themselves will be stored in a file structure at the EGDI server, from where they will be accessible for the end users. The documents (the machine-readable ones) will also be searchable by the search system.

2 TECHNOLOGY

2.1 EGDI metadata catalogue – MIcKA

The EGDI metadata catalogue uses the MIcKA system for management and publication of metadata on structured data. MIcKA technology enables entry, editing, harvesting, discovery, and view of metadata on geological data across Europe. It provides tools for compilation and export of the metadata in a standardized format.

2.1.1 Application MIcKA version 6

MIcKA is a web application for management and cataloguing of spatial metadata. It is based on the following technologies: PHP Nette framework, XLST, PostgreSQL fulltext search, JQuery, Bootstrap and OpenLayers.
For some past European projects such as OneGeology-Europe, Minerals4EU and ProSUM, the technology of MIcKA version 5 was applied to develop their metadata catalogues. This version, modified and amended, was implemented as a part of the EGDI prototype in 2016. It was also used to compile and manage metadata for the GIP project from the project beginning. In order to better meet the needs of the GIP project, the upgraded MIcKA version 6 was developed and put into operation at the end of October 2019.

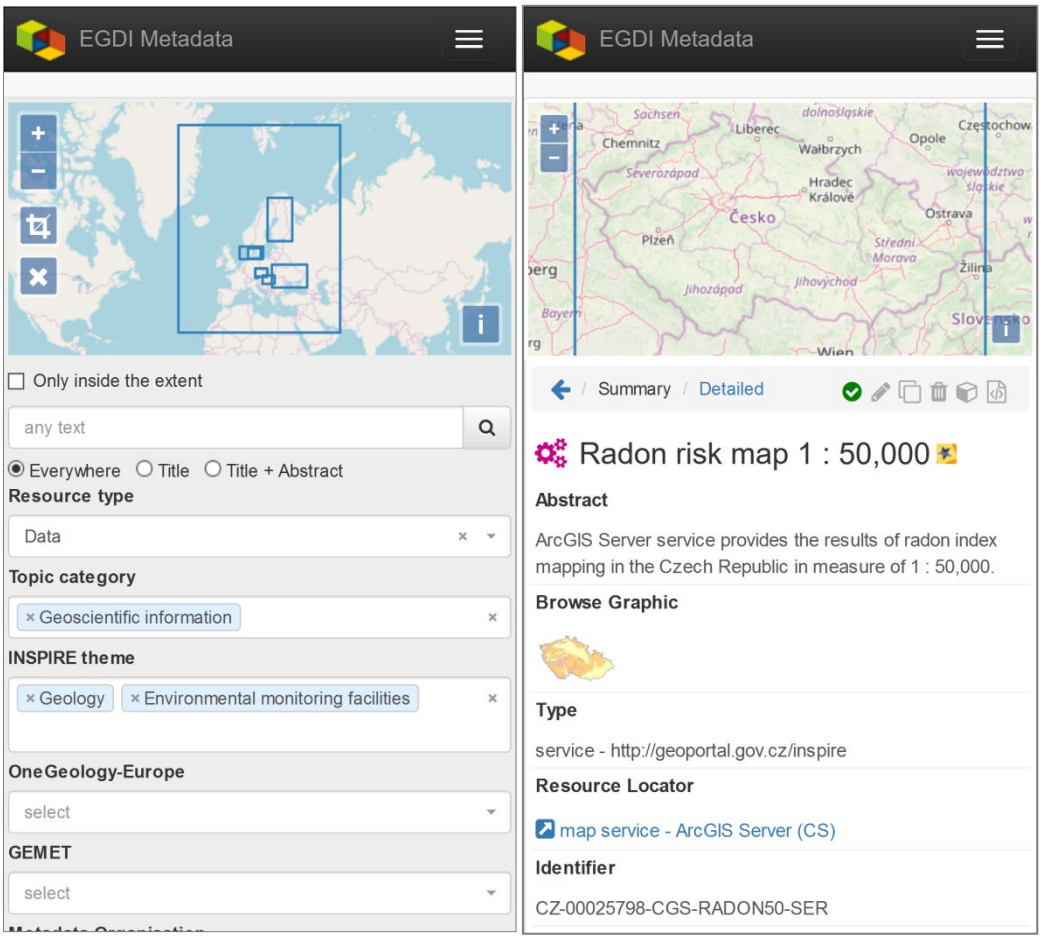

Fig. 1: Responsive design of MIcKA v.6

The new version of the catalogue includes a lot of additions and improvements in the **user interface**, such as a new and responsive design (Fig. 1), improved relation between metadata records, better themes and profile configurability, the possibility to have multiple values in search boxes enabled (Fig. 2). Also, the lite editor has been rewritten to fulfill INSPIRE 2.x profile requirements. The list of supported standards was extended (see chapter 2.1.4) and INSPIRE Metadata profile 2.x, ATOM and WFS download services **validator** was added. MIcKA also has a number of **technical improvements** (see chapter 2.1.5). The code of version 6 is available as Open Source (https://github.com/hsrs-cz/Micka).

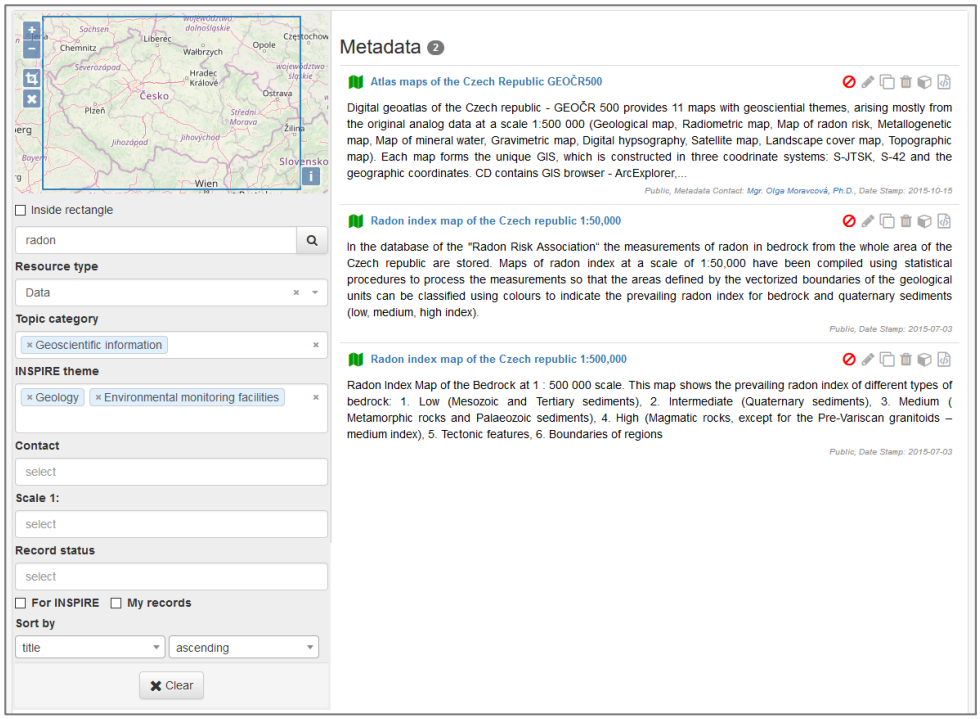

Fig. 2: MIcKA v.6 search interface

2.1.2 Database

The data is stored in a dedicated PostgreSQL database that is backed up regularly once a week. The latest version of the backup is kept for a month.

2.1.3 Operation of the system

The system runs on the operating system SuSE Linux, operating on encrypted communication https://egdi.geology.cz/ during harvesting. After any major intervention on the server, a complete server backup is created on the virtualization platform level to enable swift recovery if needed. This backup is used as a clone of the server for testing of major application changes. Regular database increments and changes in the application code are backed up once a week. In addition to that, one backup copy for each month is stored for 12 months.

2.1.4 Supported standards

The EGDI metadata catalogue supports the distributed system of metadata administration and is fully compliant with the following international standards:

- ISO 19115 Geographic Information: Metadata
- ISO 19119 Geographic Information: Services
- ISO/TS 19139:2007 Geographic Metadata XML (gmd) encoding and XML Schema implementation derived from ISO 19115
- ISO 19110 Geographic Information: Methodology for feature cataloguing
- ISO 15836:2009 - Information and documentation - The Dublin Core metadata element set
- Open Geospatial Consortium Catalogue Service for Web 2.0.2 ISO AP 1.0
- INSPIRE Metadata profile 2.x (gmx:Anchors support and corresponding codelists)

2.1.5 Input and output formats

Available input data formats are:

- ISO 19139 XML
- ISO 19110 XML (for Feature Catalogues)
- OGC WMS, WFS, WCS, SOS and CSW capabilities documents

- KML, ATOM files

Output data formats:
- ISO 19139 XML / OGC CSW 2.0.2 GetRecords/GetRecordById XMLs
- HTML with embedded RDFa or JSON-LD
- GeoDCAT-AP
- OAI-MARC, MARC21
- JSON, KML
- ATOM + INSPIRE ATOM Download service implementation

2.1.6 Technical features

Basic features:
- Web application
- Multilingual interface (and multilingual metadata editing)
- Modular, extensible
- Access to INSPIRE registry, GEMET and other thesauri, including those that will be developed by the GIP project
- Built-in validator (INSPIRE or National profiles)
- User defined profiles
- Two editing forms - full / simple (lite), configurable
- Extensible full text search
- OGC CSW 2.0.2 ISO AP 1.0 compatible
- INSPIRE metadata 2.x compatible
- Extensions (queries, outputs…)

New in version 6:
- Responsive design
- Unlimited output size for any format (useful for KML and DCAT feeds)
- Better performance
- Rewritten ordering to speed up output
- Full text search support (with linguistic support)
- Registry client implementation (INSPIRE registry, SPARQL endpoint - CGI, GEMET, etc.)

3 EGDI METADATA PROCESS SCHEMA

3.1 Metadata of the structured data

Metadata are freely accessible to the public for viewing and searching, but inserting, editing and deleting is only available for authorized users to ensure that only data from authorized organizations are being added to the EGDI infrastructure. The login and harvesting information may be obtained on request from the administrator (**egdi.metadata@geology.cz**).

3.1.1 Metadata entry

There are two ways to create a new record:

Automatically by harvesting (performed by the EGDI metadata administrator)

- One-time harvesting that moves metadata records from a temporary repository to the EGDI metadata catalogue
- Regular harvesting copies metadata from a permanently updated project repository to the EGDI metadata catalogue

It is possible to harvest either all project records, or only a subset of them (based on filtering).

Manually (performed by logged-in authorized user with editing rights)

- Creation of a new record
- Import from URL (GetCapabilities) or from the file (.XML)
- Copy of the existing record using the editing tools

3.1.2 Metadata editing/update

Regularly harvested metadata are updated automatically in a mode defined in the harvesting settings (usually a 1-day period).
Metadata records created directly in the EGDI metadata catalogue can be directly edited by the users with editing rights, when they are logged in. Metadata records may have two different record statuses: private and public (Fig. 3). Public status means that everyone can search and view the record; private records are visible only for the owner of the record.

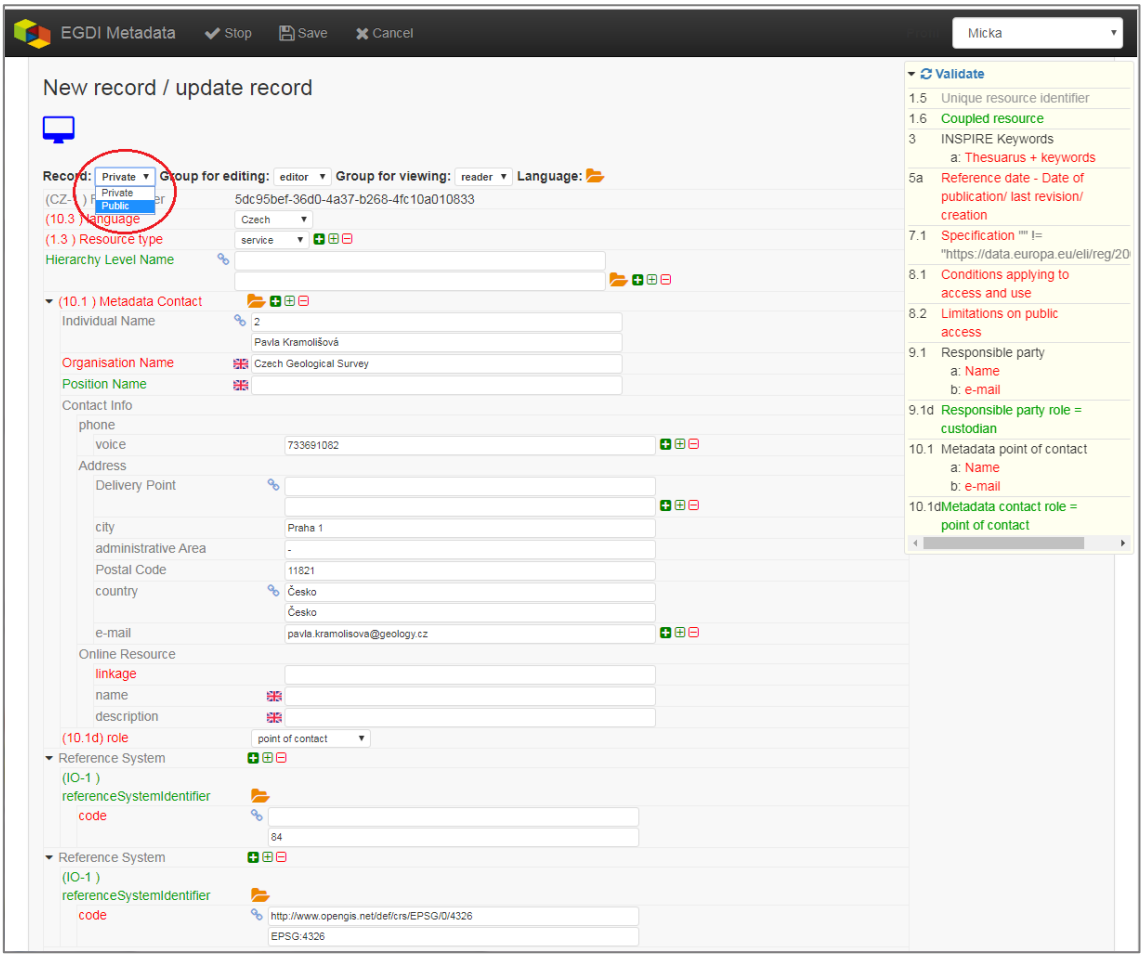

Fig. 3: Editing form of a metadata record (a record status selection tool is highlighted)

3.1.3 Metadata deletion

The harvested metadata records have to be deleted in the source catalogue – then their status will be compared during the next harvesting run to the EGDI catalogue and they will be deleted there as well. Metadata record created directly in the EGDI metadata catalogue may be deleted from there by the authorized metadata editor. Most of the GeoERA metadata records may be used on the portal for describing datasets or services and it always must be carefully checked before the deletion, so not to break a valid link.

3.1.4 Cookbook and Helpdesk for metadata creation in the EGDI metadata catalogue

For more information on the creation and use of the EGDI metadata records, a Cookbook is being created within the WP8 and will be made available accordingly. Currently its draft is placed on https://egdi.geology.cz/catalog/micka/cookbook for logged users only, publicly available help is on https://egdi.geology.cz/help.
The full version of the Cookbook and training materials will be available on https://micka-docs.readthedocs.io/en/latest/ and included in the GeoERA documentation website https://geoera-gip-docs.readthedocs.io/en/latest/. For further details, it is possible to visit https://github.com/GeoEra-GIP/Project-Support-WP8 or use the email helpdesk with any questions/issues: **support@geoera.eu**.

4 EGDI METADATA PRINCIPLES

4.1 Metadata usage

The EGDI metadata provides very important information on:

- Documentation of the dataset itself (abstract, custodian contact, data origin, keywords, time-space description, update frequency, formats, etc.)
- Documentation of the data access options (availability of data resources, possible formats and ways of distribution, links to web services or user-oriented applications etc.)

The MIcKA system as a tool for editing and displaying metadata offers especially these functionalities:

- Integrated search tools for MIcKA users (full text search, search by country, organization, keywords, harvesting source, geographical scope, etc.)
- API for third-party search tools (e. g. EGDI portal)

As MIcKA can offer output (search results) not only in HTML, but in JSON format as well, this JSON output can be used in the future for:

- Guideposts/lists generated directly from metadata (a unified data navigation system based on metadata keywords) on portal
- Display of details of a single metadata record on portal

The JSON file generated by MIcKA can be loaded, parsed and displayed on any web page by various tools – both on the server or client side. This gives a variety of options that can be applied to the Portal to improve the integration between metadata and the map viewer and enhance the user experience with the EGDI platform.

4.2 Metadata requirements

To integrate metadata effectively into the EGDI portal, the main principles must be fulfilled as following:

- All data is described by metadata in the EGDI metadata catalogue
- All metadata records have to be tagged by keywords from controlled vocabularies and thesauri as defined by GeoERA projects and GIP-P WP4
- All distribution formats are described in a standardized form in metadata to allow cross-linking between data, services, and applications
- Each data resource, including those originated from a finished project, has an active GSO´s metadata contact for maintenance
- In order to display a metadata record for which an on-line map service is available, metadata must be properly linked to the EGDI Portal
- No service appears on the EGDI portal unless it has a properly filled metadata record – ideally this should be controlled by the portal automatic monitoring system
- Each web service operates on one or more datasets. One-way relation from service to dataset is mapped with the *operatesOn* metadata element, providing link to the ISO 19139 dataset metadata record according to INSPIRE rules
- If present in the metadata catalogue, the backward mapping from dataset to corresponding services shall also be available. Mapping from a harmonized dataset to national resources (services or datasets) is provided by *dataQualityInfo/*/lineage/*/source/*/citation/*/xlink:href* element, which contains URL to the corresponding ISO 19139 metadata. The relations between services and harmonized datasets are mapped in the same way as relations between national services and datasets (Fig. 4)

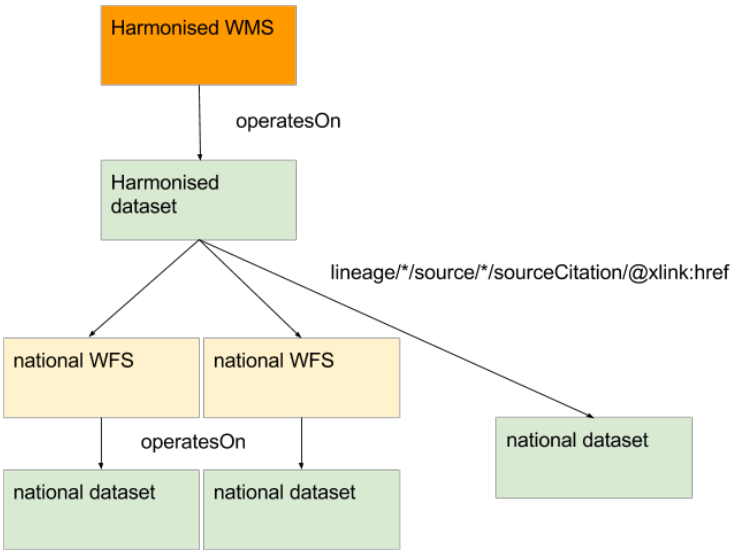

Fig. 4: Example of relations between metadata records of harmonized and national data sources

5 SCOPE OF THE EGDI METADATA CATALOGUE

5.1 Types of data to be included in the metadata catalogue

Only **digital and structured** data sources are described by metadata in the EGDI Metadata catalogue as listed below:

- Datasets
- Dataset series
- Spatial data services
    - INSPIRE view (Web Map Services - WMS))
    - INSPIRE download (Web Feature Services - WFS), Web Coverage Service and ATOM download service
- Multidimensional models
- Applications (the user-oriented access to data)

5.2 Multilinguality

The bilingual principle (English and national language) is an option widely used in national catalogues. It is also a rule for EGDI. More languages may be included.

5.3 Thesauri – keywords

The metadata catalogue is prepared for incorporation of the Common vocabulary/registry/thesaurus of geological terms as developed by WP4. **The Multilingual Thesaurus and Project Dictionaries** will be linked to the EGDI metadata catalogue when the final version will be accepted and placed (proposed data.geoscience.earth domain owned by EGS). These will be used as keywords source for metadata records, following the SKOS rules. The keywords are to be provided as URIs using gmx:Anchor element.

Existing thesauri such as INSPIRE theme, OneGeology-Europe, Gemet and some others are linked to the EGDI metadata catalogue as well.

6 EGDI METADATA PROFILE

6.1 EGDI metadata profile

The EGDI metadata profiles for structured data sources descriptions are based on the INSPIRE requirements binding in the EU Member States on the basis of Commission Regulations 1205/2008/EC, 1089/2010/EC and the related Technical Guidelines to Regulation 1205/2008 version 2.0 https://inspire.ec.europa.eu/documents/inspire-metadata-regulation and Technical Guidelines to 1089/2010 (https://inspire.ec.europa.eu/id/document/tg/metadata-iso19139). The INSPIRE requirements are further extended to include the necessary items from international standards, especially ISO 19115/19119/19139 standards.

6.1.1 GeoERA metadata profile for structured data

The metadata profile for the description of results of the GeoERA projects will be based on the EGDI profile. To enable an effective filtering, there will be requirements to tag each record with the defined keywords (project name, project vocabularies, and keywords from the geoscientific thesaurus) – these vocabularies will be defined in the upcoming period and will be incorporated in the MIcKA environment. All GeoERA projects were asked to define requirements for any possible extensions to the GeoERA metadata profile as soon as possible, so that the changes could have been defined according to existing standards, implemented and tested. As of December 2019, only requirements from the HotLime project for the description of 3D models have been obtained.

6.1.2 GeoERA metadata profile for the 3D (xD) models

The proposed first version of the metadata profile for the description of 3D models is based on the EGDI profile but also includes some additional metadata elements to describe the third dimension:

- Presentation form – mandatory element filled as „modelDigital“ (from ISO 19115)
- Model version – „edition“ metadata element (from ISO 19115)
- Vertical coordinate system – „verticalDatum“ metadata element (from ISO 19115)
- Model depth extent – „verticalExtent“ metadata element (minimum and maximum value, from ISO 19115)
- Specific keywords of types stratum, temporal, and discipline to be filled with keywords from controlled vocabularies in the form of URI (from ISO 19115)

More information and help will be made available directly in the EGDI Metadata Catalogue on https://egdi.geology.cz/help, https://egdi.geology.cz/catalog/micka/cookbook and included in GeoERA documentation website https://micka-docs.readthedocs.io/en/latest/. The full documentation will be available in 2020.

6.2 Validation

Built-in metadata validator is part of the EGDI metadata catalogue to ensure compatibility of the metadata records with INSPIRE 2.0.1 (https://github.com/inspire-eu-validation/metadata). It may be modified to include a customized project metadata profile if a specific need appears (e.g. mandatory project thesaurus keywords).
The metadata validator may be used as a standalone tool for on-line validation of the user metadata before posting them to the catalogue. INSPIRE ATOM and WFS download services validation is added as well.

7 CONCLUSIONS

A **new version of the EGDI metadata catalogue** based on MIcKA 6 technology is operational since 24.10.2019. This is the main result described by D7.1. This is one of the key parts of the EGDI infrastructure, amended and customized to better serve the GIP-P needs. It can be linked to the EGDI portal to provide information on origin of each of the data presented. The interface is ready to serve to the EGDI search engine so that users will have a user-friendly way to search through all of EGDI content.

All content from the previous version of the EGDI metadata catalogue was incorporated into the new version; the harvesting mechanism for a specific project or organization catalogues was improved and set as needed. The work on entering metadata for all GeoERA projects results can start as planned.

All previously existing EGDI metadata editors' accounts have been transferred to the new version and have been assigned the same access rights as in the old version of the catalogue. For each of the GeoERA projects it is crucial to nominate one metadata editor who will be responsible for the entering of metadata for all the projects results and will be aware of the duty to keep the metadata up-to-date even after the projects are completed.

To ease the work of the metadata editors, the Cookbook is available for authorized editors in the MIcKA application, at the GitHub helpdesk https://geoera-gip-docs.readthedocs.io/en/latest/index.html, and a specific training has been offered to all who are interested.

The basic principles of the metadata catalogue are governed by the principles of INSPIRE, which, however, change in time. On 15 December 2019 new INSPIRE Technical Guidelines for metadata will come into force, which may affect the further development of the EGDI metadata catalogue. The sustainable development of the catalogue is inevitable.