HIKE GeoERA GEO-ENERGY

Hazard and Impact Knowledge for Europe

Deliverable 5.3

Final Project Data Management Implementation Report

Authors and affiliation:
Esther Hintersberger [GBA], Serge van Gessel [TNO]

E-mail of lead authors:
Esther.Hintersberger@geologie.ac.at
Serge.vangessel@tno.nl

Version: 2021.10.06

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. Scientific work is co-funded by the Geological Surveys and national funds allocated for science within the period 2018-2021.

Deliverable Data		
Deliverable number	D5.3	
Dissemination level	Public	
Deliverable name	Final Project Data Management Implementation report	
Work package	WP5, Information Platform Interface	
WP Lead /Deliverable beneficiary	[GBA]/[GBA]	
Deliverable status		
Submitted (Author(s))	06/10/2021	E. Hintersberger, S. van Gessel
Reviewed/Verified(WP leader)	18/10/2021	R. van Ede [TNO], J. Tolstrup [GEUS] / E.Hintersberger [GBA]
Approved (Project leader)	18/10/2021	Hans Doornenbal [TNO]

TABLE OF CONTENTS

1 INTRODUCTION

1.1 Document Background and Scope

This document presents the final implementation of the HIKE Data Management Plan (DMP) that was presented during the initial phase of the project.

The purpose of the DMP is to plan measures which maximise the access and reuse of the data that is collected and generated in the context of the HIKE project for future projects and end users. These measures should help to make the data easily findable, accessible, interoperable and reusable (FAIR).

This report presents the state of knowledge about the data and metadata that the project partners have collected and generated in HIKE. The DMP addresses how the data is made accessible and exploitable for verification and re-use, and how they shall be maintained for future use. The DMP thus reflects the main elements of the data policy that the consortium has adopted regarding the data sets that have been collected and generated during the project implementation.

The HIKE databases and products have been established in close interaction with the GIP project to guarantee consistency of data management and documentation within the overarching GeoERA project and to execute the parts of the DMP that are related to the EGDI Information Platform. Interaction with the GIP has been organized by GBA under WP5 ”Information Platform Interface”. GBA and TNO are responsible for communicating HIKE information platform requirements to GIP (specifically the GIP-WP2 Liaison) and conversely ensuring that the guidelines and standards provided by GIP are properly implemented in the research proper outcomes. This has been mainly archieved by personal contact.

1.2 Documents and sources mentioned in the document

This document includes references to the following reports and sources produced by HIKE and the GIP project. Links to the HIKE documents are listed on the HIKE web site[1]

- HIKE D2.1b: Final Fault Data Characterization Catalogue
- HIKE D2.4: Final report on Fault Database Application and evaluation
- HIKE D2.5: The Fault Database embedded in EGDI and including collected partner data
- HIKE D4.2b: Final Scientific specifications and requirements for the hazards and impacts data share point and definitions for the Semantics Web service
- HIKE D4.3: The Knowledge Share point embedded in EGDI and including collected partner data
- HIKE D5.1a: Technical IP requirements of the Fault Database
- HIKE D5.1b: Technical IP requirements of the Knowledge Share Point
- HIKE D5.2b: Final user manual for the Fault Database and the KSP

[1] https://geoera.eu/projects/hike10/documents/

1.3 Abbreviations

HIKE = Project ”Hazards and Impacts Knowledge Europe”
GIP = Project ”Geo-Information Platform”
EGS = EuroGeoSurveys organization
EGDI = European Geo Data Information Platform
DMP = Data Management Plan
FDB = Fault database
KSP = Knowledge Sharepoint
SHARE = Project ”Seismic Hazards Research Europe”
EPOS = Project ”European Plate Observing System”
MICA = Project ”Mineral Intelligence Capacity Analysis”
DOI = Digital Object Identifier
URI = Unique Resource Identifier
GSO = Geological Survey Organization
INSPIRE = Infrastructure for Spatial Information in Europe
GEOSCIML = data model and data transfer standard for geological data
SI = International System of Units

1.4 HIKE partners

#	Participant Legal Name	Institution	Country
1	Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek TNO	TNO (coordinator)	Netherlands
2	Albanian Geological Survey	AGS	Albania
3	Geologische Bundesanstalt	GBA	Austria
4	Royal Belgian Institute of Natural Sciences – Geological Survey of Belgium	RBINS-GSB	Belgium
5	Geological Survey of Denmark and Greenland	GEUS	Denmark
6	Bureau de Recherches Géologiques et Minières	BRGM	France
7	Bundesanstalt für Geowissenschaften und Rohstoffe	BGR	Germany
8	Landesamt für Bergbau, Geologie und Rohstoffe Brandenburg	LBGR	Germany
9	Landesamt für Geologie und Bergwesen Sachsen-Anhalt	LAGB	Germany
10	Bayerisches Landesamt für Umwelt	LfU	Germany
11	Islenskar orkurannsoknir - Iceland GeoSurvey	ISOR	Iceland
12	Istituto Superiore per la Protezione e la Ricerca Ambientale	ISPRA	Italy

13	Servizio Geologico, Sismico e dei Suoli della Regione Emilia-Romagna	SGSS	Italy
14	Agenzia Regionale per la Protezione Ambientale del Piemonte	ARPAP	Italy
15	Lietuvos Geologijos Tarnyba prie Aplinkos Ministerijos	LGT	Lithuania
16	Państwowy Instytut Geologiczny – Państwowy Instytut Badawczy	PIG-PIB	Poland
17	Laboratório Nacional de Energia e Geologia	LNEG	Portugal
18	Geološki zavod Slovenije	GeoZS	Slovenia
19	State Research and Development Enterprise State Information Geological Fund of Ukraine	GEOINFORM	Ukraine

2 DATA MANAGEMENT EVALUATION

The following sections provide a brief overview and evaluation of the different aspects in the Data Management Plan

2.1 Data summary

2.1.1 Purpose of the data collection/generation and its relation to the objectives of the programme

The purpose of the data collection and generation in HIKE is subdivided in two objectives:

1. To collect and harmonize data on tectonic subsurface and surface faults (i.e. fracture surfaces along which appreciable displacements have taken place) and to disseminate these data to end users for a broad range of applications ranging from geological modelling to hazard research.

2. To collect and organize case-based study results, data sources, tools and methodologies, and generic knowledge related to hazard and impact research at the geological surveys of Europe. This information is intended to guide researchers and stakeholders to relevant intelligence and suggested approaches for future assessment studies.

Objective 1 has been achieved by establishing a robust and future-proof fault database (FDB) which holds the information on faults and associated tectonic features from various sources. The database has been implemented in EGDI[2] (European Geo-Data Infrastructure) to secure sustained accessibility and enable future updates. Throughout the duration of the HIKE project, fault data has been collected from the project partners (HIKE 2.2b) and other GeoERA projects (HOTLIME[3], GEOCONNECT[3]d[4], 3DGEO-EU[5]). This has resulted in an online accessible GIS - dataset that covers most of Europe (HIKE D2.5[6]). Through the associated semantic database, the data in the FDB has been linked to external fault databases. Various applications are presented in the HIKE D2.4 report[7], which includes examples from Work Package 3 “Case studies and Methods”, other GeoERA projects and national research projects.

Objective 2 has resulted in the development of a Knowledge Share Point (KSP) which allows geological surveys and science communities to upload relevant documents and information sources and link within a semantic keyword database (HIKE D4.3 report[8]). The semantic framework supports the searching and retrieval of documents by relating

[2] http://www.europe-geology.eu/
[3] http://geoera.eu/projects/hotlime6/
[4] http://geoera.eu/projects/geoconnect3d6/
[5] http://geoera.eu/projects/3dgeo-eu/
[6] https://geoera.eu/projects/hike10/faultdatabase/
[7] https://geoera.eu/wp-content/uploads/2021/10/D2.4_HIKE_Fault_DB_Evaluation.pdf
[8] https://geoera.eu/wp-content/uploads/2021/10/D4.3_HIKE_KSP_Implementation_Report.pdf

them to common hazard and impact themes (e.g. subsidence) and defining cross-thematic relations (e.g. geothermal extraction to induced seismicity triggered by geothermal activities). HIKE partners have contributed to the definition of the vocabulary terms and the uploading of documents in the KSP. The KSP repository is intended to continuously be extended with new documents in the future.

2.1.2 Types and formats of data generated and collected

Fault Database:

In general, geological faults are represented as linear features on a 2D map. The fault geometry is represented by lines on a 2D map (vector data), whereas a fault can be represented by several lines at specific depths/horizons. Besides its geometry, the fault data incorporate specific attribute data (fault properties) and meta-data which are stored in linked tables. A semantic database (project vocabulary) helps to include unstructured fault information and to link to other information sources on the internet, e.g., publications and reference or other information pages of other fault databases.

The FDB supports data delivery by the open (industry standard) format geopackage[9] (.gpkg). The exact formats and parameters for the FDB are specified in detail in the publicly available fault data specification and characterization catalogue (HIKE D2.1 report[10]).

Knowledge Share Point:

The KSP repository holds actual documents (PDF/ZIP) or DOI links to non-spatially data consisting of case studies, reports, published methodologies, tools, etc. Accepted upload formats are described in GIP report 2.3.1[11] (chapter 7). Like the FDB, the KSP incorporates a hierarchically structured keyword database which is presented in detail in HIKE Deliverable 4.2. The exact formats and parameters for the KSP are reported in Deliverable D4.3.

2.1.3 Re-use of existing data.

The partners have been encouraged to make data available from existing national mapping and research programmes. Prior to HIKE these sources were either only locally or internally available. HIKE has provided data templates and guidelines, in order to be able to harmonize the different datasets that are provided. In addition, the project has linked data from other sources via the semantic vocabulary.

Fault database:

For the FDB, the project partners have mainly provided existing data that has either not been publicly available before or was only available as printed maps. Already published

[9] https://www.geopackage.org/

[10] https://geoera.eu/wp-content/uploads/2021/10/D2.1b_HIKE_Fault_Data_Characterization_Catalogue.pdf

[11] https://geoera.eu/wp-content/uploads/2020/04/D.2.3.2-Mapping-and-describing-the-needed-extensions-to-EGDI.pdf

fault databases have been linked via the semantic vocabulary (e.g., SHARE[12], ITHACA[13], Slovenian fault data base[14]) or fault geometries and information from local or internal databases have been reused by the respective partners. Information generated in the trans-border project GeoMOL[15] has been incorporated in the data deliveries of the Austrian and Bavarian partners GBA and LfU, respectively.

Knowledge share point:
The KSP in particular represents information that is uploaded by partners or that is available on the internet. A recommended future development is to evaluate how HIKE can link with other repositories like for example EPOS[16].

2.1.4 Data origins

Fault database:
The fault data originates from national and regional data information repositories at the partner GSO's. The fault data (FDB) has mainly originated from the HIKE project partners and other GeoERA projects that generate fault information. Besides that, HIKE has included links to fault information from other repositories that are already published on the internet via the semantic vocabulary. Examples are the SHARE database on seismogenic faults and the ITHACA Italian database on capable faults. During the project, many of the partners have updated, extended and harmonized their Fault data. In some instances new mapping activities haven been conducted. All partners have established a novel tectonic boundary classification which is included in the HIKE project vocabulary.

Knowledge share point:
The KSP incorporates the results from the HIKE methodologies and case studies as well as key reports and publications relevant for hazard and impact research at the surveys. The reports are either produced in-house or they are external publications or results from (EU) research collaboration.

2.1.5 Size of data sources delivered to EGDI

Fault database:
The FDB holds a total data pool of about 100 MB, but might potentially increase to several GBs in size, depending on the scale and extent of mapped information, inclusion of additional 3D information, etc. The database is expected to grow over time (also after the project is finished). The updates will take place at national level. EGDI will be the

[12] http://www.share-eu.org/
[13] https://www.isprambiente.gov.it/en/projects/soil-and-territory/italy-hazards-from-capable-faulting
[14] https://tectonics.geo-zs.si/
[15] https://www.geomol.eu/
[16] https://tcs.ah-epos.eu/login.html

platform where updates are represented. Description of the upload procedures and templates for data preparation are available (HIKE D2.1 and D5.2b[17] reports).

Knowledge share point:
The KSP has a fair size of up to several 10's MBs which could become several GB's over time, depending on whether report entries will be uploaded as PDF/ZIP or as reference (DOI). The KSP is intended as a repository that will grow after HIKE. There are procedures for uploading/labelling new entries (HIKE D4.3 report[18])

2.1.6 Target users and data utility

Fault database:
The data in the FDB is open for everyone to use. The metadata provides background on the origins. It is expected that the FDB will initially form a basis for geological research by the geological surveys and associated scientific community:

- Before HIKE, many surveys lacked an infrastructure to store, manage and disseminate fault data. The FDB provides a platform for existing and future fault mapping, modelling and characterization projects at the national and European level.
- The faults form the basis for the Structural Framework approach (developed by GeoConnect^{3}d) which is expected to become a basis for geological models and resource assessments in future projects.
- The HIKE project itself (i.e. WP3) and various other GeoERA projects (in particular GeoConnect^{3}D, 3DGEO-EU, HOTLIME, VOGERA[19]) incorporate examples where national and regional information on faults (as represented in the FDB) is required. An overview of such applications is provided in HIKE D2.4
- The data is available and applicable for various types of projects from European Commission services and European Agencies, Geological Surveys outside HIKE, (National) stakeholders and the general public in the addressed pilot areas, Research community, Third parties using fault information (including software tools), consultancy agencies, software developers, regulators, policy makers
- The usefulness of the FDB has been extended through the possibility to directly link information on individual faults and groups of faults to other fault databases.

Knowledge share point:
The knowledge Share Point is open to diverse target user groups including.

- Developing Community includes stakeholder groups involved in HIKE and other geological survey organizations that provide, manipulate adapt and use information of the KSP to own benefits part of their research activities. This category is represented by the Geological Surveys, nation-wide research programs, public research institutes, etc.

[17] https://geoera.eu/wp-content/uploads/2021/10/D5.2b_HIKE_User_Manual.pdf
[18] https://geoera.eu/wp-content/uploads/2021/10/D4.3_HIKE_KSP_Implementation_Report.pdf
[19] http://geoera.eu/projects/vogera1/

- Expert category includes researchers, knowledgeable person, universities, research agencies, and organizations/institutes that are not directly involved in HIKE and are looking for scientific and/or specific information. This category could also provide new information to the KSP through their scientific work and high competence in a domain/concept.
- Stakeholders as regional, national and local authorities, supervisors, operators and industry looking for case studies, general information, protocols, etc.
- Education, Public is the category including students, the general public or any other stakeholder searching for topics and specific information based on scientific criteria and provided by legal and thrust sources.

2.2 Fair data

2.2.1 Making data findable, inclusion provisions for metadata

Data identification and discoverability

All data generated by the HIKE project are associated with metadata records. The FDB metadata record consists of a parent general metadata record and one child record per delivering partner. All these records are stored on the EGDI Metadata Catalogue[20]. The document entries in the KSP have their own metadata records. Both the FDB and KSP make use of vocabularies[21] and thesauri[22] for which each of the concepts, definitions and keywords are defined using URI's (Unique Resource Identifiers[23]). Referenced sources in the KSP are using a DOI's (Digital Object Identifier[24])

Naming conventions

The FDB and KSP have both defined strict naming conventions for data objects, attribute names and values, and concepts and definitions in the project vocabularies. The HIKE project is linked to an overarching keyword database developed in EGDI (e.g. geological terms). All naming conventions have been defined in the HIKE D2.1b, D4.2[25] reports and the internal specification documents shared between HIKE and EGDI (D5.1a and D5.1b).

Data, metadata, documentation and open access.

The major part of the data, metadata and vocabularies generated in HIKE are embedded in the EGDI central platform and database(s). EGDI provides the functionalities including repositories, web-GIS systems, web services such as WMS, WFS, download services and metadata catalogue which allow the straightforward discovery and visualization of

[20] http://www.europe-geology.eu/metadata/
[21] https://geoera.eu/wp-content/uploads/2019/11/D4.3-GeoERA-Project-Vocabularies.pdf
[22] https://geoera.eu/wp-content/uploads/2019/11/D4.2-GeoERA-Keyword-Thesaurus.pdf
[23] https://www.w3.org/Addressing/URL/uri-spec.html
[24] https://www.doi.org/
[25]https://geoera.eu/wp-content/uploads/2021/10/D4.2_HIKE_KSP_Specifications_Background.pdf

the data. The HIKE project web page[26] uses the EGDI services to provide direct access to the results.

2.2.2 Making data openly accessible

2.2.2.1 Access to results.

By default, all the data and associated semantic vocabularies and metadata produced by HIKE that are stored and published via EGDI and are open, following the Creative Commons BY 4.0 licence (no limitations to public access). Few deviations are possible and are listed in the country/regional-specific metadata record. The embedding of the FDB and KSP in EGDI ensures sustained access, compliance with common and open standards, and the possibility to combine data with other sources of information (either in the EGDI webGIS, through downloading and import in in-house software or via WFS/WMS services). The reports and final deliverables are available via the HIKE project web page.

2.2.2.2 Required software

Data delivery from partners and storage of the FDB on the EGDI platform use the open file standard geopackage (.gpkg), which can be downloaded and imported by commercial and open-source GIS software. A view of the HIKE FDB via the EGDI data portal, the KSP starting page, the semantic vocabulary and the metadata via the EGDI metadata portal are all accessible via any web browser application.

Web services such as WMS, WFS are standard functionalities of the EGDI platform and will allow discovery and representation of data through standardized GIS applications.

HIKE does not disseminate software, as none was developed as part of the project. The KSP source code is integrated in the EGDI platform and not available as stand-alone software. The (javascript) code could be requested.

Specifics regarding software tools and source code used in EGDI are reported via the GIP project web page[27]

2.2.2.3 Data storage and documentation

The major part of the data, metadata, reports and other documentation are disseminated via the EGDI portal which is intended to be sustained by the GSOs after the end of the GeoERA. This will ensure accessibility after the HIKE project lifetime. Reports are stored in a GeoERA repository and disseminated via the HIKE project web page.

2.2.3 Making data interoperable

2.2.3.1 Data interoperability, exchange and re-use

The EGDI platform supports combined visualization of datasets in the webGIS portal. Vocabularies allow the interlinkage of HIKE data with other repositories (specifically the FDB data with SHARE and ITHACA). Due to the storage in the open source format

[26] https://geoera.eu/projects/hike10/
[27] https://geoera.eu/projects/gip-p/

geopackage, the data in the FDB are downloadable and re-usable beyond the EGDI platform (e.g. import in in-house software or via WFS/WMS services). As all data is provided following the Creative Commons BY 4.0 licence and show no limitations to public access. What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

FDB data follow international standards such as INSPIRE, GeoSciML/ CGI, SI, and others used in EGDI, where possible. Deviation from or extension to existing INSPIRE and GeoSciML codelists are documented both in the FDB specifications (Deliverable D2.1) and in the respective linked vocabulary entries. Even though the FDB is not so, a basic INSPIRE-compliant dataset for ShearDisplacementStructures can be easily derived from the provided information. Metadata records follow the standards of the EGDI Metadata Catalogue, which should guarantee interoperability between such portals and the dissemination of project data.

2.2.3.2 Use of standard vocabularies for inter-disciplinary interoperability

In HIKE, standard vocabularies were used, where possible. We also generated a new vocabulary of named fault/tectonic boundary objects in Europe. Within GeoERA, this was partly harmonized with similar vocabularies provided by the projects Geoconnect³d and HotLime. Where possible, the vocabulary entries were linked to external sources.

2.2.3.3 Use of uncommon or project-generated ontologies and vocabularies

In the cases where HIKE deviates from or requires extension to the existing standard vocabularies, the new terms were semantically linked to the commonly used standards vocabularies (CGI and INSPIRE), if possible.

2.2.4 Increase data re-use (through clarifying licences)

2.2.4.1 How will the data be licensed to permit the widest re-use possible?

By default, all the data produced by HIKE that are stored and published via EGDI are open, following the Creative Commons BY 4.0 licence and show no limitations to public access. Few deviations are possible and are listed in the country/regional-specific metadata record.

2.2.4.2 Data re-use, embargos and patents

As published under the Creative Commons BY 4.0 licence, the data produced and published in HIKE are usable by third parties.

There are no reasons to assume a delay to release of the generated data or - in the future – the updated data. No issues are foreseen with regards to re-use restrictions or patents

2.2.4.3 How long is it intended that the data remains re-usable?

The technical durability depends on the persistence of EGDI, which is the responsibility of GIP or a follow-up. In addition, the consortium partners have local versions of their data in national or regional repositories that are maintained by the consortium partners, which are for internal usage only.

HIKE delivered fault data and hazard/impact information according to the state-of-art available from the various partners, sources and case studies at the time of the project. The coverage and detail of fault & hazard/impact information varies between countries. Depending on the progress of mapping and characterization after GeoERA, the data may be outdated less or more rapidly after HIKE has been concluded.

HIKE has established a data structure for the FDB and KSP that allows for updates after the conclusion of HIKE. However, due to the complexity of the data, the update procedure will be most likely require assistance from the maintainer of the EGDI platform. The updating of local databases after HIKE has been concluded, occurs under the responsibility of the data owner/provider. HIKE cannot guarantee or be held responsible for (timely) post-project updates.

2.2.4.4 Are data quality assurance processes described?

The FDB data collection included the following data quality control procedures (supported by reported guidelines and specifications reports):

1) Partners (GSO's) are responsible for the geological validity of the data provided. This depends on QC procedures executed at GSO level
2) The submitted geopackages have been checked on inconsistencies. All data has been mapped to code lists and is thus harmonized. TNO and GBA have performed the QC procedures on the contents of the data set (e.g. deviations from code lists, inconsistencies, etc.)
3) Major corrections have been reported back to data providers for processing and approval.

The above cycle has been repeated until both the data provider and FDB development team approved the quality and validity of the data.

After the HIKE project has ended, it will be possible to upload new data or updates. In this case the HIKE project can no longer perform the QC procedures. It is recommended that the QC steps 1-3 are followed by uploading partner and by using provided specifications and guidelines (HIKE D2.1 and D5.2b reports). At this stage there is no provision for QC step 2. It is recommended that this will be resolved within the EGDI platform.

2.3 Allocation of resources

2.3.1 Costs of data FAIR development?

One of the aims of HIKE consists in the integration of the produced data into EGDI which is managed by GIP. The project took care that the data and supplementary materials have sufficiently rich metadata and a unique and persistent identifiers to make them Findable. The Accessibility, Interoperability and Reusability are assured by the fact that all the data and metadata are available in a standardized way through the central database of EGDI and associated vocabularies. Licensing and traceability issues of the data were also taking into consideration. The centralisation of the storage and the publication of the data and metadata reduces the global cost.

2.3.2 Resources for long term preservation

The costs for the maintenance of the central EGDI platform is under consideration and beyond the scope of HIKE.

2.4 Data security

2.4.1 Provisions for data recovery and secure storage and transfer of sensitive data

This aspect is governed by EGDI

2.4.2 Certified repositories for long term preservation and curation

This aspect is governed by EGDI

2.5 Other issues

2.5.1 National/funder/sectorial/departmental procedures for data management

The underlying data from the GSOs or partners are governed by national/regional or institutional rules.