AGEMERA

Critical Raw Materials for a Resilient Europe

DATA MANAGEMENT REPORT 1

Deliverable 7.4

This project has received funding under the European Union's Horizon Europe research and innovation programme under grant agreement No 101058178.

Project no.	**101058178**
Project acronym:	AGEMERA
Project title:	**Agile Exploration and Geo-Modelling for European Critical Raw Materials**
Call:	HORIZON-CL4-2021-RESILIENCE-01
Start date of project:	01.08.2022
Duration:	**36 months**
Deliverable title:	D7.4 DATA MANAGEMENT REPORT 1
Due date of deliverable:	31.01.2024 (M18)
Actual date of submission:	31.01.2024
Deliverable Lead Partner:	**University of Oulu, Kerttu Saalasti Institute**
Dissemination level:	Public

Name	Organisation
Jari Joutsenvaara	University of Oulu, Kerttu Saalasti Institute
Eija-Riitta Niinikoski	University of Oulu, Kerttu Saalasti Institute

Document History			
Version	**Date**	**Note**	**Revised by**
01	18.1.2024	Draft version	JJ
02	30.1.2024	Draft for commenting	JJ, E-RN
03	31.1.2024	Final version	JJ

This project has received funding under the European Union's Horizon Europe research and innovation programme under grant agreement No 101058178.

Disclaimer

The content of the publication herein is the sole responsibility of the publishers, and it does not necessarily represent the views expressed by the European Commission or its services.

While AGEMERA is funded by the European Union, views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the European Research Executive Agency (REA) can be held responsible for them.

While the information contained in the documents is believed to be accurate, the authors(s) or any other participant in the AGEMERA consortium make no warranty of any kind with regard to this material, including, but not limited to the implied warranties of merchantability and fitness for a particular purpose.

Neither the AGEMERA Consortium nor any of its members, their officers, employees or agents shall be responsible or liable in negligence or otherwise howsoever in respect of any inaccuracy or omission herein. Without derogating from the generality of the foregoing, neither the AGEMERA Consortium nor any of its members, their officers, employees or agents shall be liable for any direct or indirect or consequential loss or damage caused by or arising from any information advice or inaccuracy or omission herein.

This project has received funding under the European Union's Horizon Europe research and innovation programme under grant agreement No 101058178.

Executive Summary

Data management

This report, Deliverable 7.4 (D7.4) of the AGEMERA project, describes the project´s data management activities until the M18 of the project. The report summarises the different data management, data inputs, data classification as well as data formats.

The AGEMERA project implements a combination of multisource geoscientific data and data fusion and processing powered by data analytics, data fusion and machine-learning algorithms. The data is collected and gathered into the AGEMERA platform, which combines the public, e.g. earth observation data, project-wise collected new geoscientific field data, drone, muographic and ambient seismic tomography data, and converts the information into actionable intelligence for mineral potential and mineral exploration data. The project surveys local communities' concerns and hopes regarding mineral exploration and compiles the generalised and anonymised results into an open-access database.

Organised into seven Work Packages (WPs), AGEMERA spans from assessing CRM potentials to data management and public engagement. This report outlines data types, acquisition methods, storage, and security measures, emphasising the role of data management in ensuring compliance, transparency, and long-term research impact. Detailed within are the handling of confidential data and the storage of open documents on Zenodo, the project website. EGDI platform will be used later in the project.

This project has received funding under the European Union's Horizon Europe research and innovation programme under grant agreement No 101058178.

Table of Contents

This project has received funding under the European Union's Horizon Europe research and innovation programme under grant agreement No 101058178.

List of Acronyms

AGEMERA	Agile Exploration and Geo-Modelling for European Critical Raw Materials
AGEMERA platform	The AGEMERA platform enables the efficient utilisation of multisource heterogeneous and multi-modal Earth Observation data. It serves as a data repository for the visualisation of EOD as well as 2D and 3D innovative geophysical surveys during the lifecycle of the project
AI	Artificial Intelligence
CRM	Critical Raw Material
D	Deliverable
DEM	Digital Elevation Model
DMP	Data Management Plan
GIS	Geographic Information System
SPSS	Silicon PhotoMultiplier
WP	Work Package

List of Figures

List of Tables

This project has received funding under the European Union's Horizon Europe research and innovation programme under grant agreement No 101058178.

1. Introduction

This report is the AGEMERA project report D7.4. Data Management report. It relates to the deliverable D7.3 Data management plan published at the beginning of the project. The document intends to list the data types used in the project, as well as the data acquisition procedures, data storage and data security, among others.

Data management is an integral part of Horizon Europe projects, ensuring compliance, transparency, and the long-term impact of research. Effective data management practices benefit not only the projects themselves but also the broader research community and society by enabling the sharing of knowledge, innovation, and solutions to pressing challenges. This document outlines data types, acquisition methods, storage, and security measures, emphasising the role of data management in ensuring data security, compliance, transparency and long-term research impact. Detailed within are the handling of confidential data and the storage of open documents on Zenodo, the project website.

In this report, data management actions during the first period (M1-M18) are described. The document also includes a section on data protection concerning activities that involve the local public and stakeholders in WP2 and WP5.

This project has received funding under the European Union's Horizon Europe research and innovation programme under grant agreement No 101058178.

2. Data

The project generates and re-uses a combination of multisource geoscientific data, newly generated data from new innovative exploration methods, field studies, and online and local surveys on social sustainability and responsibility. (see Figure 1. for AGEMERA data process flow). There are many kinds of data in the project: text documents, PowerPoint presentations, Excel sheets, raw measurement data, processed data, SoftGIS data, UAVs, and satellite data.

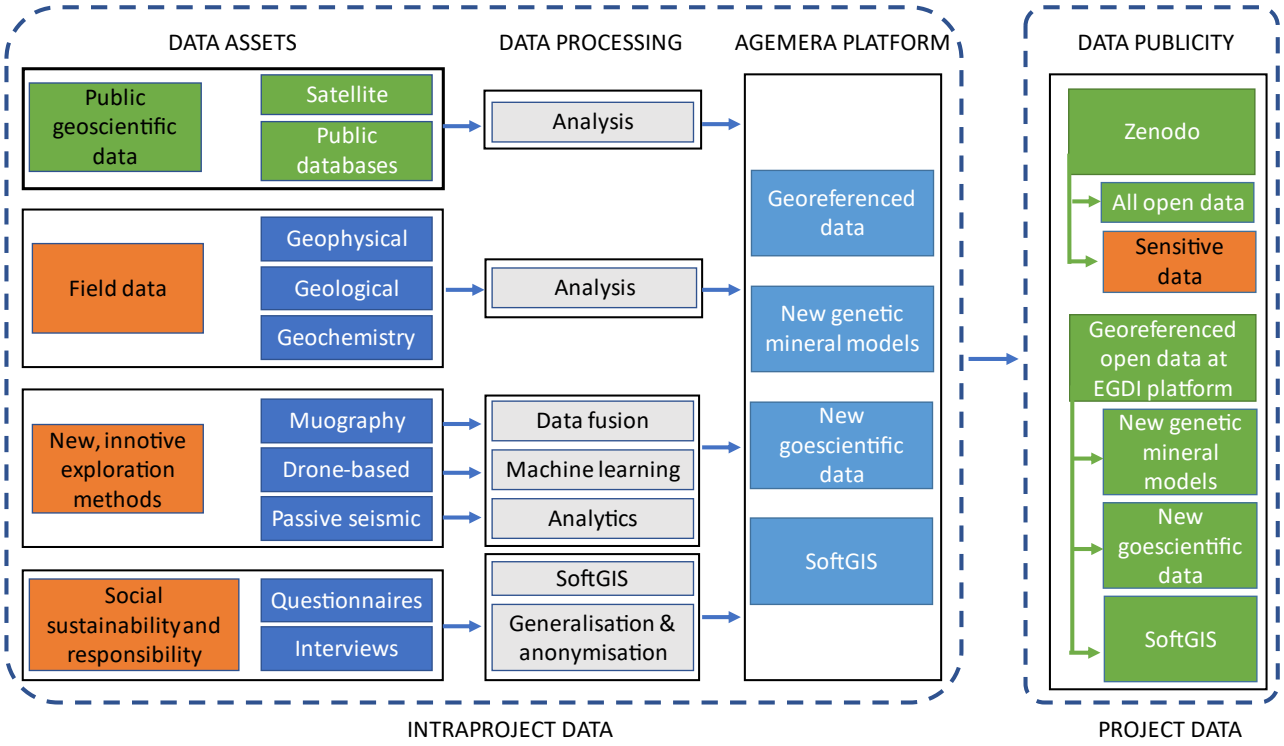

Figure 1. Primary data sources for the AGEMERA project. More detailed descriptions and explanations of the data and process are in the text.

In Figure 1, the colours define the data publicity: green = public, orange = sensitive. Data assets marked with blue are partner-owned data, which use can be sensitive and partner-restricted. Data processing methods are marked with grey. Light blue describes the processed data, which is either public or sensitive. The data publicity column describes the plans for publishing the public data sets and making them available for further use. Field data is collected project-wise, and public results will be published. Technology developers' data is sensitive, as is social data. Processed public data will be published on Zenodo and the EGDI platform.

This project has received funding under the European Union's Horizon Europe research and innovation programme under grant agreement No 101058178.

2.1 Data formats

In the AGEMERA project, many different data sets have been collected, from social science and project management to geophysical and geological. The used data formats include such as ASCII, CSV, JSON, GEOTIFF, COGTIFF, JPG, shapefile, NetCDF, miniSEED, SAC, GRD, DXF/DWG, RD and SPSS formats; they are explained briefly in the following text.

The ASCII (American Standard Code for Information Interchange) is a character encoding standard used for representing text in computers and other devices that use text. Due to its lack of metadata and limited character set, ASCII is typically used for simple text data. CSV (Comma-Separated Values) is a file format used for storing tabular data (numbers and text) in plain text form. While CSV is a widely used format for its simplicity and readability, its lack of inherent metadata means that external information is often necessary to understand and utilise the data it contains fully.

JSON (JavaScript Object Notation) is a lightweight data-interchange format that is easy for humans to read and write and easy for machines to parse and generate. JSON is text-based and made up of objects and arrays. JSON files often contain implicit metadata in the form of the keys used in key-value pairs. While it can inherently carry some level of metadata through its key-value pairs, complex data structures may require additional schema definitions or external documentation for full context and validation.

DXF (Drawing Exchange Format) and DWG (Drawing) are two common file formats used in computer-aided design (CAD). DXF is a text file that contains a representation of the CAD drawing. DXF files are useful for sharing vector graphics or designs across different platforms and software. DWG is a proprietary, binary file format. It stores design data, metadata, and user data. DWG files are more compact and complex than DXF and are typically used for storing and sharing CAD drawings

NetCDF (Network Common Data Form) is a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data. NetCDF is widely used in the atmospheric and oceanic sciences and other fields. There are two common file formats associated with NetCDF: .nc and .cdl. The .nc is the binary file format used for efficient storage and access of multi-dimensional scientific data in NetCDF, while .cdl is a text-based format used to describe the schema of NetCDF files in a human-readable form.

JPEG (Joint Photographic Experts Group) is a commonly used method of lossy compression for digital images, particularly for those images produced by digital photography. JPEG files can contain metadata such as camera settings, GPS data, and other information embedded within the file, usually in the EXIF (Exchangeable Image File Format) format. JPEG's balance between compression efficiency and image quality has made it a standard in digital photography and online image sharing. However, for applications requiring lossless image storage, other formats like PNG or TIFF are generally preferred.

This project has received funding under the European Union's Horizon Europe research and innovation programme under grant agreement No 101058178.

A Shapefile is a popular geospatial vector data format used in GIS (Geographic Information Systems) software. It stores the location, shape, and attributes of geographic features, like points, lines, and polygons. A Shapefile is actually a collection of several files. The main file, .shp, stores the geometry of the features. The .shx is an index file that facilitates access to the features in the .shp file. The dBASE table, dbf, contains feature attributes in a tabular format.

Regarding metadata, a Shapefile does not inherently include extensive metadata about the data itself (like coordinate system or data creation information). However, it is common to accompany a Shapefile with a separate metadata file (often in XML format) that provides this additional context. This metadata file is not a standard component of the Shapefile format itself but serves as a crucial supplement for fully understanding and using the Shapefile data.

GeoTIFF format allows the georeferenced metadata to be stored in the metadata of the TIFF image. TIFF stands for Tagged Image File Format and is a commonly used raster graphic file type. The digital elevation models (DEMs) are an example of information stored in GeoTIFF files. The GeoTIFF associates the evaluation (z) for every coordinate (x,y), and it also includes info on the used Coordinate Reference System (CRS). This allows the representation of pixel coordinates in real-world coordinates. Cloud Optimised GeoTIFF (COG) is an enhanced version of the standard GeoTIFF file format. COG is optimised for web use, enabling efficient access to geospatial raster data stored in cloud environments. This makes it particularly useful for large datasets commonly used in remote sensing and geographic information systems (GIS). COG is especially beneficial for satellite or aerial imagery, digital elevation models (DEMs), and other large raster datasets frequently used in Earth observation and spatial data analysis.

The miniSEED and SAC are two prominent file formats used in the field of seismology, including ambient seismic studies. miniSEED (miniature SEED) is a subset of the SEED (Standard for the Exchange of Earthquake Data) format and is primarily used for time series data in seismology. It includes very limited metadata beyond basic time series identification and simple state-of-health flags. It does not encompass detailed metadata such as geographic coordinates or response/scaling information, which are crucial to interpret the data values. The SAC (Seismic Analysis Code) format is optimised for data analysis, although it is more challenging for storage. This format is defined by the SAC software but is supported by various other seismological tools. Besides the waveform data, SAC files can include metadata that describes critical information about the waveform, such as station and earthquake information, instrument response, and more.

The RD file format is commonly used in X-ray diffraction (XRD) analysis. They typically contain data like the intensity of diffracted X-rays and the angles at which diffraction occurs. RD files are used for further analysis and interpretation of the crystalline structure of materials.

This project has received funding under the European Union's Horizon Europe research and innovation programme under grant agreement No 101058178.

SEG-2 is a data file format used primarily in geophysical applications, particularly for seismic data. It is designed to store seismic reflection, refraction, and other related data. A key feature of SEG-2 is that it includes metadata, which provides crucial context for the seismic data. This metadata often includes information about the seismic source, geophone layout, recording parameters, and other details essential for interpreting the seismic data accurately.

The GRD file format, often used in geophysical, geological, and geographic applications, represents grid data. This data typically consists of information like gravity intensity, magnetic readings, or colour, structured as a 2D array on a plane. While the format efficiently handles spatial data, it does not inherently contain detailed metadata within the file itself. However, certain software applications working with GRD files can extend their capabilities to include or associate metadata, such as coordinate system information, through external or companion files.

Maps are used and stored as georeferenced formats such as shape files. Social scientific data is stored as SPSS (Statistical Package for the Social Sciences), which uses a proprietary file format for storing data. The typical file format used by SPSS is the .sav file. It stores both data and metadata (like variable definitions, missing value definitions, etc.). The .sav format is binary, which makes it efficient for storage and fast to read and write. This file format can handle a wide range of data types, including numeric, string, and date variables. In a typical .sav file, the data is structured in a tabular format similar to a spreadsheet. Each row represents a case or an observation, and each column represents a variable.

2.2 Geoscientific data

The geoscientific data includes geological, geophysical, geochemical, and physical samples from the field and analyses of the results of these. This data has been collected in WP1 and WP3 and processed in WP4. The drone-based surveys produce electromagnetic, magnetic and radiometric data. The data produced is in ASCII and corresponding geophysical characteristic maps. The muon-based density data is in ASCII and NetCDF formats. The ambient seismic data deals with many physical parameters and thus utilises multiple data formats such as miniSEED, SAC, SEG2, ASCII and GRD.

Field and reference data from study areas are in multiple formats, such as GeoTIFF, TIFF, JPG, CSV, and DXF/DWG. These data sets represent geological, geochemical, geophysical and structural data.

This project has received funding under the European Union's Horizon Europe research and innovation programme under grant agreement No 101058178.

2.3 SoftGIS data

SoftGIS refers to a type of Geographic Information System (GIS) that integrates 'soft' geospatial data, primarily derived from people's perceptions, experiences, preferences, and feelings about places. In WP2, softGIS data has been collected through a questionnaire on people's perception of mineral exploration and mining in their home regions. The format of this final, sensitive data is in SPSS file formats and maps.

2.4 Other data

Project Management Data: This includes documents in standard Office formats like Word, PowerPoint, and Excel, along with their templates. These files are essential for project documentation and are stored in the intraproject repository. Project-related visual image versions are stored in JPG format.

Geospatial Data: Derivatives of open-source satellite imagery from sources like Sentinel-1, Sentinel-2, and ASTER have been collected for field trials and study areas. These data sets are formatted in COGTIFF, a specialised format for georeferenced raster imagery, enhancing their utility in geospatial analysis and GIS applications.

This project has received funding under the European Union's Horizon Europe research and innovation programme under grant agreement No 101058178.

3. Data management

In the project, the collected data and access to data have been controlled. In Table 1, there is an overview of data, different data sets, and the structure of how the data management was done. All the data comes with ownership and description. The data storing is organised as illustrated in Figure 2.

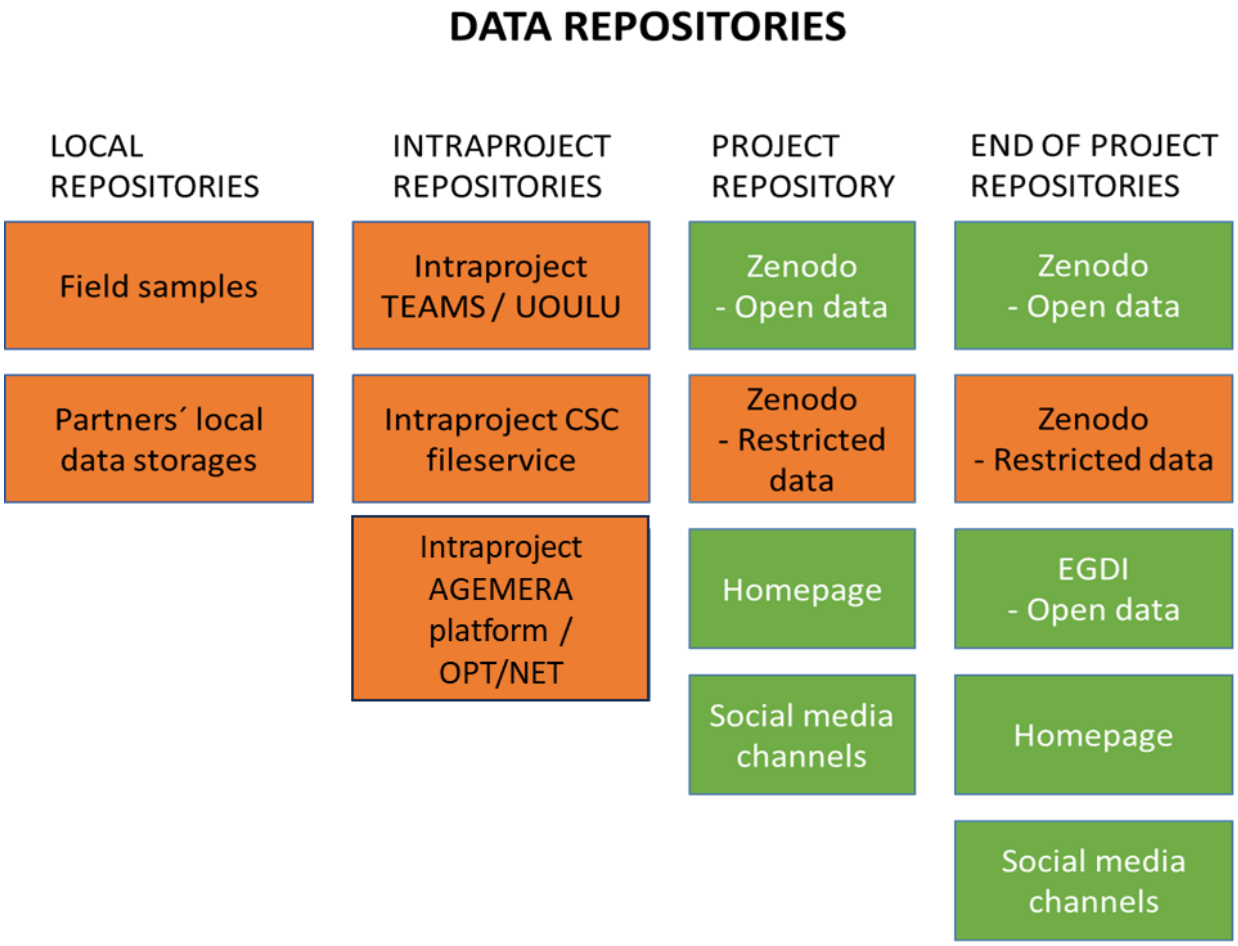

Figure 2. Schematic view of AGEMERA project´s data repositories.

Various repositories are used throughout the AGEMERA project. Local repositories are used to store generated data, and intraproject repositories are used to share data among the project consortium. Project repositories are used to store final datasets and promote the project and project outcomes. End of the Project repositories are used to give access to the project's public data. Orange-labelled data, marked in Figure 2, has restricted access due to sensitive data, and green-labelled data is open data.

The main data platform has been the project's Microsoft Teams workspace, with workspaces for all the work packages and project management. Access to the workspace is only available to the consortium. Sensitive data, such as those created in questionnaires in WP2, is secured in local repositories with controlled access only to

This project has received funding under the European Union's Horizon Europe research and innovation programme under grant agreement No 101058178.

those working on the data. Partners working on technological development use their own local data repositories.

Table 1. Example excerpt of the data sets and data types summary.

WP	Data description	Public / Sensitive	Data type	DOI or other identifier	Ownership
2	Questionnaire data	Sen	SPSS	X	X
1	New EDGI CRMs data for Bulgaria and the defined polygone	Pu	Excel	https://doi.org/10.52215/rev.bgs.2023.84.3.133	BAS
1	Petrographic photographs Ni-Co Pyrite belt	Sen	JPG	X	X
3	Muon-based density survey data	Sen	Ascii	X	X
3	Magnetometry - filtered data - 11 profiles	Sen	Ascii	X	X
4	Data from rock strength measurement on rock samples	Sen	xls	X	X
4	Derivatives of open-source (Sentinel-1, Sentinel-2, ASTER) satellite imagery for field trials = clustering analysis, RGB composites, change detection, spectral indices	Pu	COGTiffs	No	OPT
7	Project management data	Sen	Office	X	X

The AGEMERA platform enables the efficient utilisation of multisource heterogeneous and multi-modal Earth Observation data. It serves as a data repository for the visualisation of EOD as well as 2D and 3D innovative geophysical surveys during the lifecycle of the project. In the AGEMERA platform, the data is accessible to the registered consortium members, but data access can be restricted if the source data is sensitive.

This project has received funding under the European Union's Horizon Europe research and innovation programme under grant agreement No 101058178.

The Coordinator of the AGEMERA project keeps track of the data sets and data types collected and managed during the project. Table 1. is an example excerpt of the data sets and data types summary. Each data is described with sensitive status (public/ sensitive), data type, possible identifiers and metadata, if applicable, and ownership of the data set. Table 1. is just an excerpt of the collected data summary. Ownership data for sensitive data has been retracted.

This project has received funding under the European Union's Horizon Europe research and innovation programme under grant agreement No 101058178.

4. Conclusions

The AGEMERA project has utilised data from various sources, including project-based geoscientific data, softGIS and public earth observation data. It has been vital from the beginning of the project to define the file formats specifically so that the collected data is compatible with the project users and for the developed AGEMERA platform so that the data can be read without errors and use the raw data for the project applications.

Depending on the data type and confidentiality, the data has been stored in many places, including Teams, the Zenodo repository, and the AGEMERA platform. After the project, the open data and documents are preserved in the Zenodo repository at www.Zenodo.org, and the geoscientific data is also preserved at the EGDI platform at https://www.europe-geology.eu/.

This project has received funding under the European Union's Horizon Europe research and innovation programme under grant agreement No 101058178.