scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A Smart Data Hub to integrate subsurface properties into model-based decision-support tools

Qian Chen[1] ✉, Nino Menzel[2], Marc S. Boxberg[2] & Julia Kowalski[1] ✉

The selection of suitable nuclear waste disposal sites is a long and challenging process, involving multiple technical and environmental assessments. For data-integrated simulation models to function as credible decision-support tools in this sensitive context, their data sources must meet rigorous transparency and reproducibility standards. This paper introduces a Smart Data Hub that provides the reproducible data foundation required to address the limits found in existing research, which rarely features both a dataset containing uncertainty information and an effective way of application for specific use cases. The Smart Data Hub is an integrated solution consisting of two main components: a dataset compiled from 50 literature sources covering geological information, structural data, and rock properties for potential repository sites in Germany, and a functional module supporting effective assembly of data compilations for a given geological structure. This approach provides reliable, reproducible data compilation for model-based decision support in the site selection process by providing transparent, uncertainty-informed data coupled with intelligent assembly capabilities.

Background & Summary

After decades of using nuclear power commercially, not much of it remains as Germany is transitioning into using green energy[1]. One aspect that has remained and will remain so for thousands or even millions of years is high-level radioactive waste. Currently, the only solution is to dispose it underground for the foreseeable future until it becomes harmless. Although the process is defined in the Standortauswahlgesetz (StandAG), or Site Selection Act[2], finding the right location is difficult. The right place entails successfully integrating multiple geological factors, including very low rock permeability, sufficient thickness and depth, sufficient area for repository construction, and demonstrated viability over long-term timescales as specified in the Endlagersicherheitsanforderungsverordnung, or Final Storage Safety Ordinance[3]. Given the number of variables present, setting up testing sites to test their long-term performance is not feasible. This fact, however, moves the spotlight onto complementary simulations, which in turn are used to narrow down possible sites by providing a detailed outlook on the future potential of each individual area, e.g., simulation-based assessment of radioactive contamination through groundwater flow. These simulations enable an assessment of site suitability prior to committing any noteworthy amount of monetary or resource investments, making them a key component of the data-driven decision-making process. Naturally, the outcome of such a simulation is only as good as its input—the data provided.

This data has to enable effective assembly into compilations that match specific geological structures and geographic areas, which demands an effective tagging system capable of filtering and organizing information while preserving data integrity throughout the process. The task becomes more complex when taking into account variations in rock properties depending on location and geological age, creating uncertainties that directly influence how nuclear waste interacts with its environment over extended periods of time. Simply acknowledging the existence of these uncertainties is not sufficient. Instead, they must be quantified and integrated into analytical frameworks[4]. Moreover, given the long-term consequences of repository site selection, every simulation must abide by principles of transparency and reproducibility, enabling independent verification by third parties. To sum up, four focal points arise: First, data collection from existing sources or dedicated campaigns must be comprehensive and extend beyond a single source for each needed parameter, even if this leads to redundant or even

[1]Methods for Model-based Development in Computational Engineering, RWTH Aachen University, Aachen, Germany. [2]Geophysical Imaging and Monitoring, RWTH Aachen University, Aachen, Germany. ✉e-mail: chen@mbd.rwth-aachen.de; kowalski@mbd.rwth-aachen.de

overdetermined data to describe a certain rock type. A direct consequence of this is that there might be inconsistencies and conflicts in the provided information, say, various permeability values per rock type. The second focal point, hence, is to provide strategies to address this challenge, either by considering data configurations associated with a certain model setup based on a manually selected subset of the data or by designing intelligent compilation strategies. The benefit of a comprehensive data point is that it contains information on variability within a parameter of interest, facilitating explicit forward and backward uncertainty quantification, which constitutes our third focal point. Finally, the fourth focal point ensures that the data-integrated workflow functions as part of a model-based decision-support tool that enforces a high level of transparency and reproducibility standards. At a more general level, information management is tackled by the FAIR Guiding Principles, a framework for scientific data management across multiple domains. According to the guidelines, datasets have to meet the following four guiding principles: Findability, Accessibility, Interoperability, and Reusability, to ensure transparency, reproducibility and reusability[5]. A long-standing example of such a metadata dataset is the GFZ Data Services by the GFZ Helmholtz Centre for Geosciences, which has been running since 2004 and strictly abides by the FAIR principles. While not specific to radioactive waste disposal, it demonstrates the practical maturity of such an implementation[6]. An international example by the Nuclear Energy Agency is the Radioactive Waste Repository Metadata Management initiative[7], launched in 2014 to harmonize metadata across 12 nuclear waste management organizations, and its successor, the Working Party on Information, Data and Knowledge Management, which expanded this work into a holistic framework for Information, Data and Knowledge Management for radioactive waste management and decommissioning[8]. At a European programme level, the European Joint Programme on Radioactive Waste Management (EURAD), running from 2019 to 2024, included multiple work packages on knowledge management with a specific goal of creating a common framework to structure knowledge[9]. This framework would then be picked up afterward by EURAD-2, which already published a formal data management plan[10]. While the above-mentioned initiatives laid the groundwork in terms of frameworks and common vocabularies, the translation into operational, site-specific datasets requires addressing practical constraints and the means to overcome them, which are the focus of this work.

When addressing practical constraints, those related to data are most significant. Because, while substantial data exists, it is characterized by significant fragmentation due to the distribution of datasets across multiple agencies, each collecting data for diverse projects. An example of how multiple agencies with overlapping research can cause inefficiencies and information scattering can be seen in the following. In the federal state of Baden-Württemberg, which hosts possible repository locations, the Landesamt für Geologie, Rohstoffe und Bergbau, or State Office for Geology, Raw Materials, and Mining, provides valuable geological maps for the state[11]. These maps, however, are artificially constrained by political borders despite the transboundary nature of geological formations. The Bundesanstalt für Geowissenschaften und Rohstoffe (BGR), or Federal Institute for Geosciences and Natural Resources, provides federal-level coordination and centralized data in some instances, as is the case with the creation of a map containing drilling locations in Germany[12]. Besides governmental institutions, important providers of geological information are research institutions. A noteworthy example is the geophysical information system FIS Geophysics by the Leipzig Institute for Applied Science[13], serving as a dataset containing petrophysical and borehole information from Germany. Nevertheless, current data management demonstrates deficiencies regarding effective data compilation strategies and the capturing of uncertainties. This issue further transcends into research where data may be compiled, but important metainformation, including detailed descriptions, involved agencies, or location, is too often not recorded[14–17]. Consequently, researchers face numerous obstacles when considering the fragmented data landscape. The first challenge is data availability, and when it is available, it is scattered across a diverse number of institutions, raising barriers to efficient data compilation for simulations. The process gets only further complicated by a plethora of licensing policies. Moreover, current datasets lack an adequate structure to support simulations required for site evaluation in accordance with minimum repository requirements. Simultaneously, only limited uncertainty information is provided. The Smart Data Hub (SDH), in turn, cuts down on these constraints, with data contained in the SDH being published under CC BY license, enabling efficient compilation of data for specific simulations while incorporating uncertainty quantification.

Figure 1 provides an overview of the SDH architecture with its formulated requirements and external points of interaction. The overarching SDH consists of two core components: a dataset and a processing layer. The dataset was compiled from 50 different literature sources[14–63], systematically organized by geographic subareas with their corresponding host rocks. Each regional entry contains detailed information on density, porosity, hydraulic conductivity, intrinsic permeability, specific heat capacity, thermal conductivity, P-wave velocity, S-wave velocity, and electrical resistivity, accompanied by uncertainty quantification based on rock type and age. Increasing the SDH functionality includes methods for tagging and merging of data, as well as sensible defaults. The tagging method enables users to effectively manage metadata by filtering relevant data and consolidating information for specific simulations. The merging process allows aggregation of properties from different sources while accounting for data uncertainties and gaps. Whenever data is missing, sensible defaults can fill the gap, ensuring data integrity.

While the research primarily pertains to Germany, its implications extend globally. The value of reusable data remains particularly valuable across Europe as geological formations span political borders. Prime examples are Switzerland's Mont Terri Underground Research Laboratory and the Benken borehole, both able to provide data relevant to the Opalinus Clay area extending into southern Germany. This, in turn, means that data in our dataset serve broader European geological applications.

The SDH demonstrates inherent versatility, extending its applicability beyond nuclear waste disposal. For other fields, especially those related to transitioning towards more sustainability, such as CO2 storage, geothermal energy, and groundwater management, the SDH design proves equally valuable. This versatility greatly increases the value of the system and encourages joint development through diverse user communities.

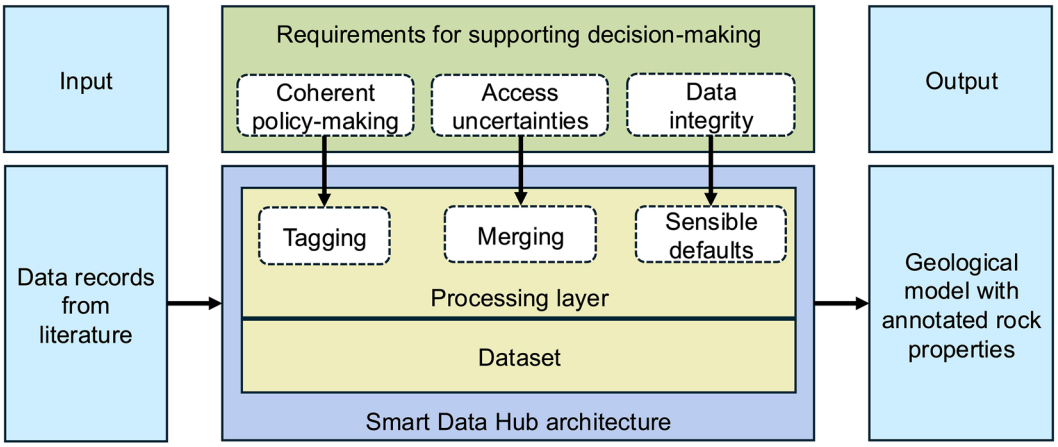


Fig. 1 Smart Data Hub structure with formulated requirements and external points of interaction. The SDH transforms scattered geological datasets as input into tailored simulation outputs. It does so through its two-layered architecture: 1) Dataset: Compiled from literature sources, with region- and rock-type-specific petrophysical properties. 2) Processing layer: Implementation of tagging for coherent data management, merging for data integration with account for uncertainties, and sensible defaults to ensure data integrity.

In summary, the SDH collects and centralizes data and tags simulations according to the data involved and contained in the SDH. This systematic approach addresses limiting factors in this field, enabling efficient data compilation for specific simulations while incorporating uncertainty quantification. Increasing transparency and reproducibility are key whenever a data-integrated simulation is applied as a basis for decision-making. By leveraging the SDH, independent third parties can reproduce results from previous simulations while enjoying the benefits of a single, easily accessible data hub. As Germany continues tackling nuclear waste management challenges, the SDH provides a solid foundation for integrating data describing subsurface properties into reliable model-based decision-support tools, enabling policymakers to better account for safety and environmental considerations when deciding on a final repository location. Given the systemic approach of the SDH, its scope of application reaches far beyond Germany. Immediate beneficiaries of the SDH in this context may be countries in close proximity, e.g., those facing a similar issue and sharing similar geological features. With the input data adjustable by the end user, the output will enable a wide range of possible applications, both geographic and geological.

Methods

This section outlines the framework created for systematic geological data compilation and integration for nuclear waste repository site assessment. This approach tackles current limitations in terms of data accessibility and fragmentation through leveraging comprehensive compilation and processing methodologies.

Data Collection. Accurate prediction of the long-term behaviour and future risks of subsurface repository sites is essential for supporting decision-making processes for repository location and design. Physics-based simulations that explicitly account for uncertainties in geological and hydrothermal environments are employed to quantify potential radioactive contamination in groundwater[4,64,65]. Optimal experimental design identifies measurement configurations that maximize information gain from geophysical surveys, such as electrical resistivity tomography and seismic refraction[63,66] and monitoring boreholes, subsequently reducing predictive uncertainty. To obtain the parameter values critical to those simulations, we collect nine rock properties in total: density, porosity, hydraulic conductivity, intrinsic permeability, specific heat capacity, thermal conductivity, P-wave velocity, S-wave velocity, and electrical resistivity. Table 1 describes each property in further detail.

Hydraulic conductivity depends on both the rock and fluid properties. Therefore, during the processing of data, the hydraulic conductivity K will be converted into intrinsic permeability k, given the density of fluid ρ_f (kg m^{-3}), acceleration of gravity g (m s^{-2}) and dynamic viscosity μ (m^2 s^{-1}) according to the following formula:

$$k = \frac{K\mu}{g\rho_f} \tag{1}$$

Developing robust and highly efficient data-integrated simulation workflows requires systematic structural management of data. Therefore, the YAML format[67] was chosen for storing rock properties due to its readability and processing capabilities. As a non-proprietary format, YAML offers universal compatibility across operating systems and is further enhanced by its nested mapping style, which logically organizes geological data. Additionally, it can be converted to multiple working data formats in Python, such as dictionaries and Pandas DataFrames[68], to streamline data processing.

Data was primarily sourced from governmental reports[14–17,19,22–24,26–28,30,33,34,37–40,42–50,53–55,57] presenting results, with a focus on potential host rocks for high-level radioactive waste, related to the site selection process. As defined in the StandAG, only three types of host rocks, namely crystalline, clay, and salt rock, are eligible for

Property	Unit	Description
Density	$kg\,m^{-3}$	Bulk, dry bulk, or grain density, with bulk density being primarily collected
Porosity	—	Effective or total porosity. Effective porosity is primarily collected as interconnected pores providing active flow paths for groundwater
Hydraulic conductivity	$m\,s^{-1}$	Ability of a fluid to move through the porous rock medium
Intrinsic permeability	m^2	Quantifying the permeability of the rock medium to fluid
P-wave velocity	$m\,s^{-1}$	Velocity of seismic compressional (longitudinal) body waves
S-wave velocity	$m\,s^{-1}$	Velocity of seismic shear (transverse) body waves
Specific heat capacity	$J\,kg^{-1}\,K^{-1}$	Heat required per unit mass to increase the temperature by one unit
Thermal conductivity	$W\,m^{-1}\,K^{-1}$	The ability of rock to conduct heat
Electrical resistivity	Ωm	Electrical resistance of the rock medium

Table 1. Overview of collected rock properties in the SDH.

final repositories, consistent with the main host rocks considered by many other countries[69]. Thus, we focus on collecting data for these three host rock areas. Complementing sources are journal articles[25,29,35,36,41,51,52,59–63], books[18,20,21,31,32,58], and a dataset[56] with CC BY license to enrich the datasets and also fill gaps wherever governmental data was insufficient. For each value taken from literature, we record meta-information of the rock type, the publishing agency, the measurement location or method, and most importantly, the original data reference. The BibTeX citation key functions in this case as an intermediate identifier for the data's origin. All sources found during the data collection are stored in a BibTeX file inside the dataset, with DOIs or URLs (if DOIS are not available) provided to each reference. Data sources not available for public download[28,45–47,57] were directly requested from BGR. As a first step, the source in question needs to be located within the BGR online catalogue. The catalogue provides a unique signature for each source, which can then be used to request the needed source from the BGR library via e-mail. Data is recorded in three distinct steps.

Step 1: Candidate Site. On the candidate site level, we record corresponding lithostratigraphic information for each rock formation in a YAML file, including geological age, lithological description and simplified lithology. The YAML file also contains paths directing to the detailed rock properties and geometric data. Table 2 details the entry fields for each site.

Step 2: Geostructural Data. We record geostructural information provided by the CHRISTA-II[14], AnSichT[37,43], and KOSINA[55] projects for each candidate site. Geostructural information, i.e., interface and orientation points for individual lithological rock formations in a geomodel, is stored in CSV files. The 3D geological structural models were created using the GemPy[70] and PyVista[71] software packages based on this geostructural data. As for the defaults, universal cokriging was applied to generate the lithological interfaces[70]. These default models provided by the SDH are completely optional and can be used when the user has no model available. The 3D geological structural model provides geometric features of the subsurface, with an example in Fig. 2.

Step 3: Rock Properties. On the rock properties level, we record all properties defined in Table 1 for a single rock formation in a YAML file. For each data point we found in the literature, we record: the BibTex citation key indicating the data's origin, the data value type, data value, unit in SI, uncertainty information (mean, standard deviation, or data range), and lithological information including rock type and additional tags (agency, location, and simplified lithology). Each property features a series of entries following a block mapping structure. Table 3 provides an in-depth explanation of entries within a block mapping.

Rock Property Metadata Convention. The simplified lithology field in Tables 2 and 3 identifies the rock types constituting each rock formation. Clear rock classification with unambiguous names is needed to support the data merging process. The International Union of Geological Sciences Commission for Geoscience Information Geoscience Terminology Working Group (GTWG) has been developing the internationally accepted geoscience vocabularies. These vocabularies include lithological rock names, hierarchically organized with up to six levels[72]. For simplicity, we categorized all sedimentary rocks mainly by grain size, regardless of sedimentary material type. Please note that we do not further classify limestone, mudstone, and sandstone. For example, claystone and silicate mudstone are classified as "Mudstone". For igneous rocks, we apply a two-group classification: Crystalline igneous rocks and Fragmental igneous rocks, although the latter is not relevant for nuclear waste disposal. This classification approach is based on the Rock Classification Scheme by the British Geological Survey[73]. All fragmental igneous rocks defined in GTWG are classified as "Fragmental_Rock", while all other groups of igneous rocks are classified as "Crystalline". Lastly, non-carbonate salt rocks, mainly consisting of evaporite minerals, are classified as "Rocksalt". Please note that this simplification strategy aims to assist data integration during uncertainty analysis and to provide sensible defaults when data is unavailable.

Geological Background. Data collection revealed that geological formations spanning political borders enable the use of international data relevant to German geological conditions. The following geological background provides the geological basis for cross-border formations and establishes guidelines for incorporating international data where German data remains insufficient.

Field	Sub-Field	Sub-Sub-Field	Data Type	Description	Additional Note
name	—	—	string	A generic candidate site name	It follows the rule of {ISO 3166 code}_{region}_{host rock type} or omitting "region" where it is not applicable
description	—	—	string	A short description for the site	
geometry_input	source	—	string	Reference to the geological structural data's origin	The entry here corresponds with the BibTeX citation key found in source.bib in the SDH
geometry_input	description	—	string	A short description of the data	
geometry_input	point_data_path	—	string	Path to the point data	The point data is stored in a CSV file
geometry_input	orientation_data _path	—	string	Path to the orientation data	The orientation data is stored in a CSV file
rock formation	source	—	string	Reference to the geological information's origin	The entry here corresponds with the BibTeX citation key found in source.bib in the SDH
rock formation	geological_time	econ	string		e.g., Phanerozoic
rock formation	geological_time	era	string		e.g., Mesozoic
rock formation	geological_time	period	string		e.g., Triassic
rock formation	geological_time	epoch	string		e.g., Middle
rock formation	geological_time	age	string		e.g., Anisian
rock formation	description	—	string	A short description of the rock formation	
rock formation	simplified_lithology	—	list of strings	A list of simplified lithologies	e.g., [Mudstone, Sandstone]
rock formation	property_path	—	string	Path to the property data	
rock formation	geometry_surface _output_path	—	list of strings	A list of paths to the surface meshes	The surface meshes are stored in VTK file format

Table 2. Definition of entry fields for site-specific YAML structure. The site-specific YAML file summarizes the rock formations at the site and establishes mapping paths to the subsurface properties, including their corresponding rock formation, rock properties, and geologic information.

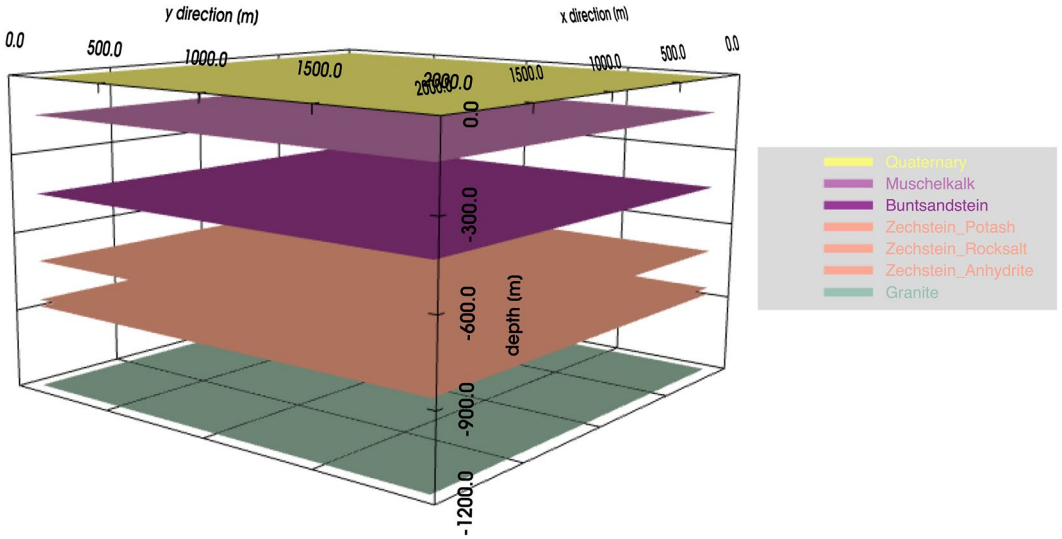


Fig. 2 A generic geological structural model for the Crystalline host rock area. Colored interface layers indicate the bottom boundaries of the rock formations. The German terms Buntsandstein and Muschelkalk refer to the lower two stratigraphic units of the German Triassic age[99], corresponding to the lower and parts of the middle Triassic according to the International Chronostratigraphic Chart[100]. The figure is generated using PyVista[71].

1. Rock salt: For Mesozoic and Palaeozoic rock salt, European deposits are predominantly characterized by extensive stratiform layers and complex salt-tectonic structures formed during the Zechstein age[74]. According to BGE (2020)[75], the most substantial rock salt formations are located within the Northern German Zechstein Basin, part of the Mid-European Permian Basin extending from eastern Britain to central

Field	Sub-Field	Data Type	Description	Additional Note
source	—	string	Reference to the data's origin	The entry here corresponds with the BibTeX citation key found in source.bib in the SDH
type	—	string	Type of the value	It can be either scalar, dictionary, or expression
value	—	float or dictionary or string	Best estimate value	If the value type is a dictionary, it is given as key/value pairs. The key refers to the "variable".If you want to enter an expression, format it as a string, for example, 5*x+7
value_min	—	float	Minimum value	
value_max	—	float	Maximum value	
value_std	—	float	Standard deviation	
probability _distribution	sample_size	int	Number of samples	
probability _distribution	sampled_data	string or numpy list or numpy array	A scipy.stats.rv_continuous().rvs() function that generate random variates or a list of custom-generated samples	e.g.,scipy.stats.norm(loc=0,scale=1).rvs (size=100,random_state=2)
unit_str	—	string	SI unit of the property	e.g., "kg/m^3" for density
unit_base	—	list of integers	The exponent of SI base units:[kg, m, s, K, A, mol, cd]	e.g., [1, -3, 0, 0, 0, 0,0] for density
variable_name	—	string	The key or function argument in the dictionary or expression	
variable_unit _str	—	string	SI unit of the variable	
variable_unit _base	—	list of integers	The exponent of SI base units: [kg, m, s, K, A, mol, cd]	
description	—	string	Detailed lithological description on the data	
tag	agency	list of strings	A list of involved agencies	e.g., [BGR]
tag	location	list of strings	Locations for data's origins	e.g., [Gorleben, Lower-Saxony, Germany]
tag	simplified_lithology	list of strings	A list of simplied lithologies	
tag	id	string	Unique id	UUID format with 36 characters[95]

Table 3. Definition of YAML entry fields for material properties. This YAML file contains detailed information about the nine rock properties described in Table 1 for each rock formation. Included are the data source, units, the description of the data value, and other metadata.

Poland[76,77]. Additionally, noteworthy Mesozoic and Palaeozoic rock salt deposits are present in the Pripyat Basin in Belarus and Ukraine, the Lublin Trough in eastern Poland, and the Cheshire and Wessex Basins in southern England[78–81]. Furthermore, regarding Triassic rock salt, despite limited studies on Triassic rock salt for nuclear waste disposal in Germany, extensive research on geothermal energy, energy storage[79,80], and large-scale geophysical surveys[82] of Triassic rock salt offers valuable insights into its geometry and geomechanics.

2. Claystone: Three major European depositional phases, Jurassic, Triassic, and Cenozoic, are relevant due to their thick, well-compacted claystone formations suitable for repository development.

   - Prevalent in Jurassic claystone, spread throughout Lower Saxony, Southern Germany, and in Southwestern Europe, including the Paris Basin and the Massif Central, is Opalinus Clay[83,84]. The Opalinus Clay formation, extensively studied by the Swiss agency NAGRA, continues from the Alpine region into Southern Germany, including the Swabian Alb and the Molasse Basin, reinforcing their importance as reference sites. Other comparable Jurassic clay-rich deposits across Europe include the Callovo-Oxfordian claystone in the Paris Basin and the Toarcian claystone in France's Massif Central[84,85].
   - The Triassic claystone in Germany exists primarily within the Keuper rock formation[37,43], the uppermost unit of the German Triassic, which is widespread across the German Basin. This basin extends from England, through northern Germany and Poland, stretching into southwestern Germany, Switzerland and parts of France[86]. Comparable deposits mainly occur in the western Molasse Basin, such as the anhydrite-bearing Keuper formations in northern Switzerland [87], and the North German Basin[43].
   - Cenozoic Claystones occur in Northern Germany and distinct regions such as the Southern German Molasse Basin and the Upper Rhine Graben[50]. The Boom Clay Formation in Belgium also represents a noteworthy Cenozoic Claystone formation. However, reduced compaction levels and relatively constrained thicknesses of up to 100 meters imply that data derived from these formations, including from the High Activity Disposal Experimental Site (HADES) underground lab, must be carefully evaluated for applicability to geological contexts in Germany[88,89].

3. Crystalline Rock: While crystalline rock is a suitable host rock for high-level radioactive waste repositories, its utilization potential in Germany remains limited due to scarcity in extent and distribution[75]. While occurring throughout the Central German Uplands and subsided zones of the Central German Upland Range, particular emphasis must be placed on Crystalline host rocks in the Saxothuringian Zone, as it contains the Mid-German Crystalline Zone, notably hosting several large granitoid bodies[14]. The Mid-German Crystalline Zone

represents a segment placed within the broader European Crystalline Zone, a geological feature stretching from the Iberian Peninsula through the Atlantic margins of Belgium and Northern France, to Central Poland and the Czech Republic[90,91]. A distinct lack of underground research facilities and data for the Saxothuringian region forces reliance on international research, such as Äspö in Sweden, Olkiluoto in Finland, Grimsel in Switzerland, and Bukov in the Czech Republic[92]. As only the Bukov underground research laboratory was situated in metamorphic rocks formed by the same Variscan orogeny that also influenced the development of relevant German crystalline bodies, well-characterized research sites comparable to crystalline rocks in Germany are strictly limited[93,94].

Data Processing. The inherent efficiency of the SDH derives from three core processing components: tagging, merging, and the leverage of sensible defaults. These components enable efficient data compilation for specific simulations while maintaining uncertainty quantification. The default data implementation ensures data integrity. Each of these components is further described in subsequent sections.

Metadata Management. Metadata management, achieved through tagging collected data, is key to tailoring specific simulation scenarios and to facilitating the reproducibility of simulations. As shown in Table 3, four tag types are added to the data: agency, location, simplified lithology, and ID. The first three tags are manually added and are defined as follows:

1. The word "agency" derives from the German words for Ministry, Society, and Office (or Ministerium, Gesellschaft, and Amt in German). If the involved party differs from the just-mentioned definition, then it is classified with the agency tag "Others". Entities falling under the "Others" tag include the following:
   - Research not done through a federal agency, e.g., universities.
   - A private company or institute hired through an agency to conduct research.
2. The "location" tag refers to the data's origin, beginning with the country and then followed by the sub-state or exact location, e.g., a city or region.
3. The previously described rock naming strategy serves as a guideline for the simplified lithology tag.

Lastly, each data point has its unique ID. It is an auto-generated UUID, namely Universally Unique IDentifier[95], based on the contents of all other entry fields. The unique ID, automatically added as a tag, enables tracking of data used in a specific simulation, further enhancing reproducibility.

Merging. With a large amount of data existing, it is not surprising that different sources provide a multitude of values for a particular property. Accounting for the multitude of values while also ensuring ease of use is difficult, especially when factors such as data inconsistency, limited availability of metainformation, and uncertainties must be considered. As for the merging process, the SDH user may choose between two options. Firstly, manually choose the tagged value for a particular property from a certain source. Secondly, merge the data employing a simple methodology, then use the output for further analysis. Due to the low porosity of dense rocks, no distinction is made between values of the same type of property when merging them, as defined in Table 1. For example, both effective porosity and total porosity are treated the same when merging. When applying the simple merging methodology, two types of information are considered for storing:

1. The respective minimum and maximum values from the given data.
2. Probability distribution of an existing dataset.

To obtain distribution information, the idea is to generate a truncated normal distribution from the merged datasets to be physically meaningful. While automated merging considers all provided sources to minimize subjectivity of the analysis, it is sometimes necessary to consider expert knowledge during data preparation. In these cases, users may opt to utilize their preferred merging method by exporting the dataset prior to applying the provided merging solution. The provided merging method consists of the following:

1. If a probability distribution is not provided, create a Probability Density Function (PDF) for each data set. The recommended PDF for each entry case is listed in Table 4. Entries not qualified are excluded from the analysis. For properties with a single value and/or a maximum range, a lognormal distribution is recommended. The standard deviation is then calculated using a recommended coefficient of variation (CV), defined as the ratio of the standard deviation to the mean. Property-specific CVs are derived from the SDH data and can be obtained from the SDH source code.
2. Generate samples from each PDF and keep only those within three standard deviations of the mean. Users can determine the number of samples to generate by entering the sample size, as described in Table 3.
3. Combine all sample sets and then fit them into a truncated normal distribution by calculating the mean, standard deviation, minimum, and maximum from the combined sample set.
4. Store the probability distribution, mean, standard deviation, and maximum and minimum values from the sample dataset.

value	value_min	value_max	value_std	Recommended PDF	Additional Note
x	x	x	x	truncated normal distribution	
x	x	x		PERT distribution	Project Evaluation and Review Technique (PERT)[98]
x		x	x	lognormal distribution	
x	x		x	lognormal distribution or truncated normal distribution	Assume a truncated normal distribution for porosity with a maximum value of one
x	x			lognormal distribution or PERT distribution	Assume a PERT distribution for porosity with a maximum value of one
x		x		lognormal distribution	Standard deviation assumed and calculated from the property-specific CV provided by the SDH
x			x	lognormal distribution	Ensure all values are greater than zero
	x	x		uniform distribution	
x				lognormal distribution	Standard deviation assumed and calculated from the property-specific CV provided by the SDH

Table 4. Recommended PDFs based on data availability. "value" denotes the best-estimate value; "value_min" and "value_max" represent the observed range; "value_std" refers to standard deviation. The "x" marker indicates a non-empty entry where data is available and recorded from the cited source. Four types of recommended distributions are provided by the SDH: 1) truncated normal distribution: Applicable if the data source provides at least the best-estimated value, minimum value, and standard deviation. 2) PERT distribution: When the best-estimated value, minimum value, and/or maximum value are available. 3) lognormal distribution: Requires at least the best-estimate value. 4) Uniform distribution: When both minimum and maximum values are available.

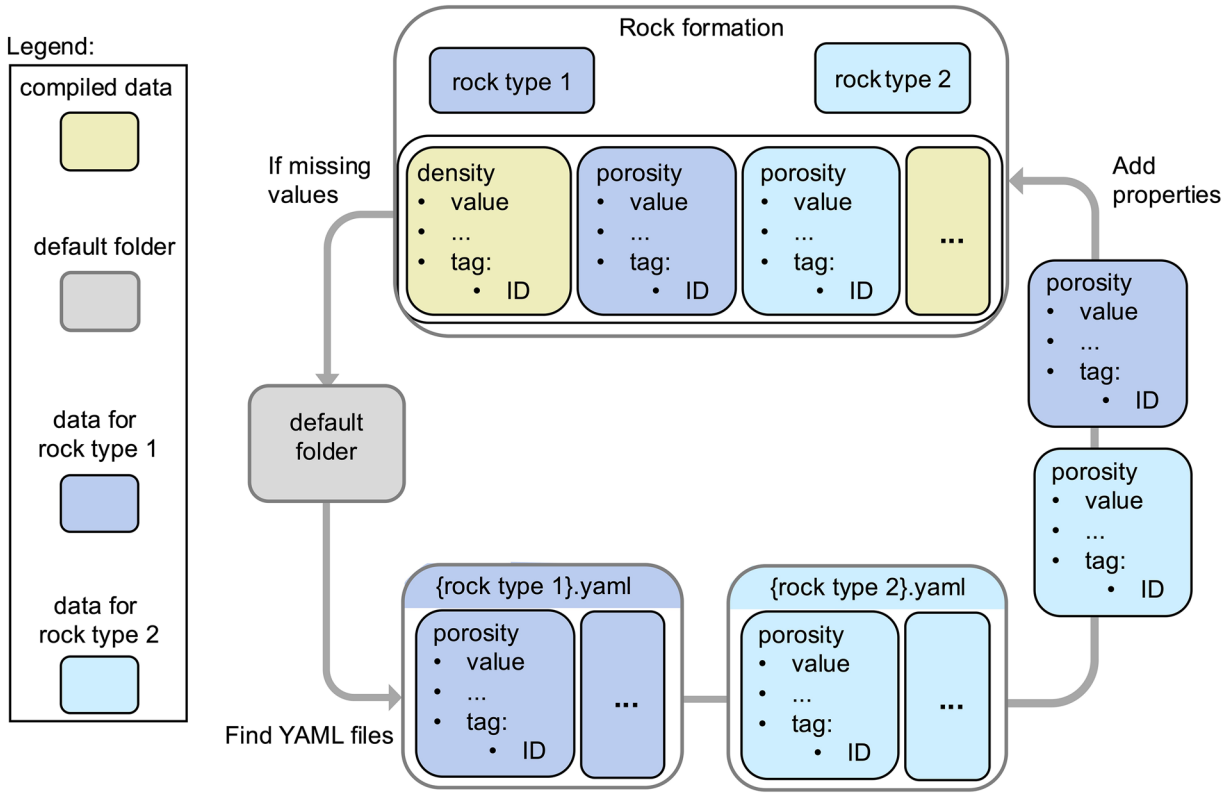


Fig. 3 Default mechanism: An example of complementing missing values. Original data records from literature sources are colored dark yellow. The colored grey is the default folder. Default data from rock type 1 is colored dark blue, and the default data from rock type 2 is colored light blue. In this example, the missing porosity value is complemented with default values based on the rock types in the rock formation, namely rock type 1 and rock type 2.

Sensible Defaults. Gaps in the data compilations naturally arise, as it would be prohibitive to restrict the SDH to those datasets that are complete. Such gaps are filled using a sensible default strategy, providing default data linked to the rock types whenever values are otherwise missing. We provide default data for the rock properties described in Table 1 and applicable to the following rock types: Conglomerate, Mudstone, Sandstone, Limestone, Dolomite, Rocksalt, and Crystalline. The default data for each rock type is calculated using the merging method

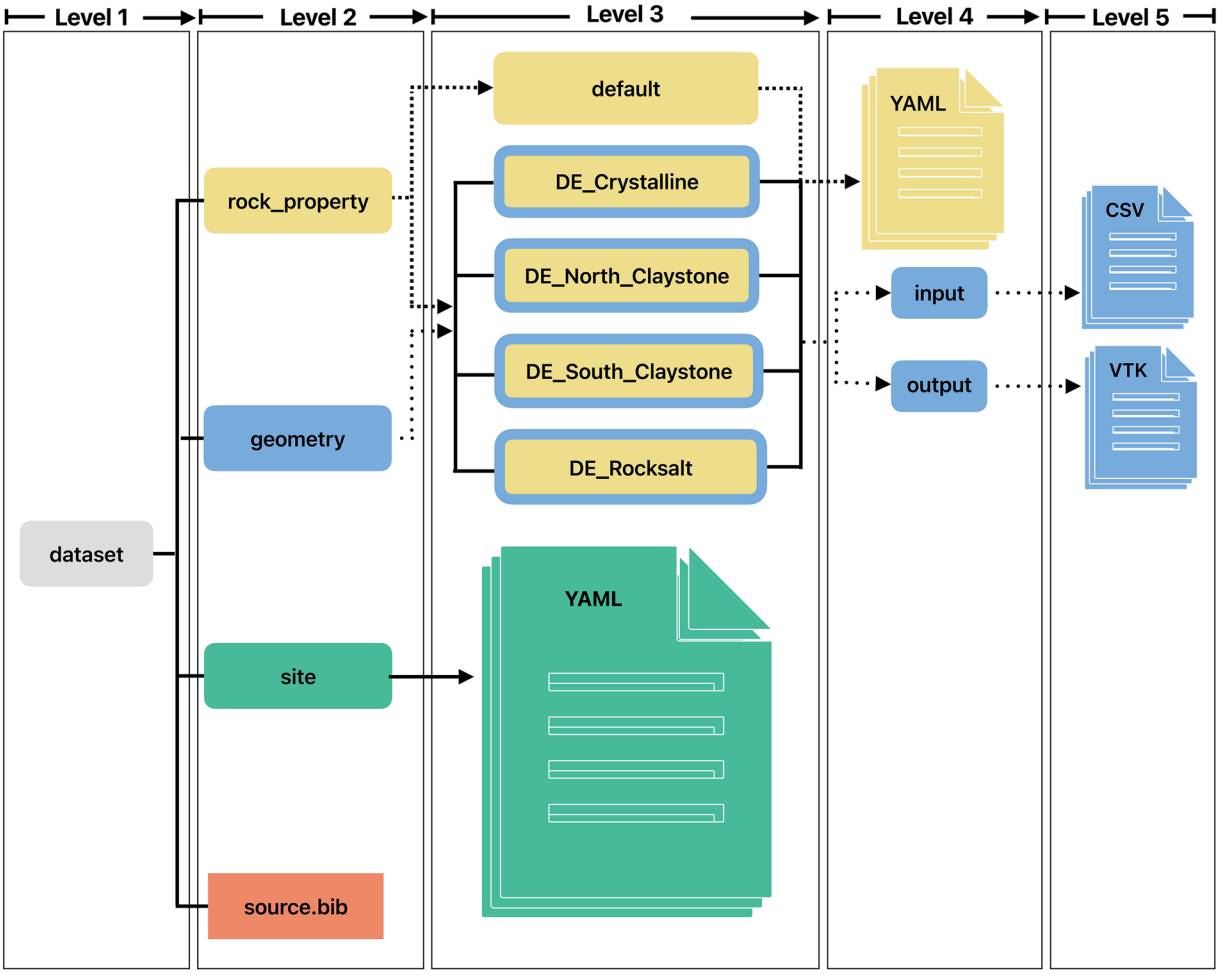


Fig. 4 Dataset folder structure. Colors differentiate folders and subfolders associated with different parent folders. Additionally, dashed lines indicate subfolders or files in the rock property folder, and dotted lines indicate them in the geometry folder. The solid line at level 3 indicates folders existing within both parent directories. For clarity, the README files are not shown.

on the existing dataset data. Encountering a scenario where default values become relevant, the user may choose between three options. Firstly, achieve completion by automatically filling in default values wherever values are missing. Secondly, provide data manually and tag them accordingly. Lastly, export an incomplete dataset. In the first case, if a rock property is missing for a specific rock formation, it will be retrieved from reference files named according to the rock types of that formation. Figure 3 further explains the process.

Data Records

The static version of the SDH dataset is available in its newest version on Zenodo[96]. It consists of three main folders, which are organized hierarchically, as shown in Fig. 4. The three folders are described below:

1. site: site-specific geological information on each lithological rock formation. It refers to a potential repository capable of containing high-level radioactive waste. Each site contains data representing a generic candidate site in Germany, including a summary of its rock formations.
2. geometry: structural information on each lithological rock. Its subfolder is further divided into two folders. The first folder contains CSV files, and the second folder contains VTK files. These CSV files are used as input to create the 3D geologic structural model, and the VTK files are output for each rock formation.
3. rock property: properties for each rock formation, including nine properties described in the section on rock properties. Contained within are the YAML files for each rock formation and its default data. The default folder contains the following YAML files: Conglomerate, Limestone, Mudstone, Sandstone, Dolomite, Rocksalt, and Crystalline.

The data comprises 50 literature sources[14–63], including peer-reviewed articles and official government reports. The source for each data value described in Table 2 and 3 is linked to a BibTeX citation key. All sources, with their corresponding BibTeX entries, are listed in the source.bib file. Figure 5 shows the distribution of data sources by location.



Fig. 5 Treemap visualizing the number of sources by location. Organized in a hierarchical structure, the treemap begins with the world, then specifies individual countries, and is followed by a sub-state or, if known, an exact location, e.g., a town or municipality. Different colours indicate different countries, while colour intensity indicates the precision of the location. The figure is generated using Plotly[101].

Field	Sub-Field	Entry
source	—	merged
type	—	scalar
value	—	5340.93
value_min	—	3125.44
value_max	—	7276.21
value_std	—	400.65
probability_distribution	sample_size	1000000
probability_distribution	sampled_data	scipy.stats.truncnorm(-5.53, 4.83, loc=5340.93, scale=400.65).rvs(size=1000000, random_state=21)
unit_str	—	m/s
unit_base	—	[0, 1, -1, 0, 0, 0, 0]
description	—	A truncated normal distribution fitted from 4 combined datasets with ids:7cd75673-f1cc-5219-bebc-d2a80a717a91, dbd050a8-58eb-5a70-b1e0-dcd26e68b295, 9133dc67-9ea4-5ea4-bfd6-62ea62d05014, e31c19ee-c7cb-5959-92f4-d31aa9be3206
tag	ID	ca63bdd4-01ca-5279-8bcc-c0e14f6ca71b

Table 5. Annotated P-wave velocity merged from data tagged with "BGR" for the Granite formation. It includes details on the proposed PDF, metadata about the original datasets, and an ID for this merged dataset.

Technical Validation

The Smart Data Hub is the tool to be used as a starting point to facilitate simulations, as it combines the key principles of reproducibility, accessibility, and transparency, all of which are essential to finding a final repository location. Ensuring the FAIR principles, the SDH serves as a strong foundation to build upon, especially when the stakes are high. Under these circumstances, informed decisions require reliable data records and the SDH outputs.

The SDH sourced data from governmental reports or peer-reviewed articles, guaranteeing its reliability. Detailed metainformation for each dataset provides the source, including the agency involved, the location, and a detailed lithological description, making everything transparent. A unique ID is provided to each source, further improving traceability. As laid out in the Methods section, tagging, merging, and sensible defaults are leveraged with a scripted programming language, making the processes involved more transparent and less prone to oversight, hence increasing overall accuracy.

Field	Sub-Field	Entry
source	—	default
type	—	scalar
value	—	2158.62
value_min	—	1178.04
value_max	—	3434.84
value_std	—	356.23
probability_distribution	sample_size	1000000
probability_distribution	sampled_data	scipy.stats.truncnorm(-2.75, 3.58, loc=2158.62, scale=356.23).rvs(size=1000000, random_state=21)
unit_str	—	m/s
unit_base	—	[0, 1, -1, 0, 0, 0, 0]
description	—	A truncated normal distribution fitted from 3 combined datasets with ids: 14eb4543-2d4e-5832-b004-095bb45c79b4, fa363cab-9c14-58c2-b20d-6ea412f6fe48, 8f3a9174-81b5-5a5e-929f-7961f9127bfc
tag	ID	d3ba3320-f314-5145-ac4e-9f7ecde2a636

Table 6. Loaded S-wave velocity from the default folder for the Granite formation tagged with "BGR". It includes details on the proposed PDF, metadata about the original datasets, and an ID for this default dataset.

The SDH systematically addresses the four focal points established at the beginning of the paper through its three core capabilities: tagging, merging, and sensible defaults. Working either independently or in combination, they enable 1) comprehensive data acquisition, followed by 2) intelligent compilation methods supporting various data configurations associated with a certain model setup, 3) explicit numerical input data uncertainty quantification and integration; and last, 4) strict enforcement of transparency and reproducibility standards.

Output from the SDH includes a 3D geological model with detailed rock property annotations, which greatly benefits from its three key capabilities. Please note that the models provided in the SDH are default geological models. However, users have the flexibility to use their own models while simultaneously accessing data from the dataset. The annotated rock properties are stored in YAML files with an example demonstrating how these capabilities work together in Tables 5 and 6.

The mean, standard deviation, maximum, and minimum values, as well as the probability distributions in Tables 5 and Fig. 6, were calculated using the data merging method. Table 5 shows the results of four merged datasets containing P-wave velocity measurements tagged with "BGR" for the Granite formation in the crystalline host rock area.

In this specific scenario, missing rock properties, including intrinsic permeability, specific heat capacity, electrical resistivity, and S-wave velocity, are sourced from the sensible defaults. Table 6 presents the S-wave velocity values loaded from the default folder.

Overall, the application of these three capabilities yields high levels of reproducibility, accessibility, and transparency. It therefore fulfils the four focal points established in the introduction: comprehensive data acquisition, customization of specific simulation scenarios, and data integrity. Furthermore, being the fourth focal point, the generated data facilitates subsequent modelling and uncertainty quantification, ultimately supporting decision-making in the nuclear waste site selection process.

Data availability

The dataset associated with this article is version v1 and is published on Zenodo (https://doi.org/10.5281/zenodo.19886769) under a CC BY 4.0 license.

Code availability

The code for processing the data, including all methods described in this paper, is available on Zenodo (https://doi.org/10.5281/zenodo.19922612)[97] and GitHub at the following link: https://github.com/mbd-rwth/smart_data_hub.git. This article corresponds to version v1.0.2.

Received: 22 December 2025; Accepted: 29 May 2026;
Published online: 23 June 2026

References

1. Bundesministerium für Wirtschaft und Klimaschutz. Energiewende und Klimaschutz: Herausforderungen und Wege der Transformation. Tech. Rep., Bundesministerium für Wirtschaft und Klimaschutz. https://www.bundeswirtschaftsministerium.de/Redaktion/DE/Publikationen/Klimaschutz/klimaschutz-energiewende.pdf?__blob=publicationFile&v=8 (2024).
2. Bundesministerium der Justiz und für Verbraucherschutz. Gesetz zur Suche und Auswahl eines Standortes für ein Endlager für hochradioaktive Abfälle (Standortauswahlgesetz - StandAG). https://www.gesetze-im-internet.de/standag_2017/BJNR107410017.html (2017).
3. Bundesministerium der Justiz und für Verbraucherschutz. Verordnung über Sicherheitsanforderungen an die Endlagerung hochradioaktiver Abfälle. https://www.gesetze-im-internet.de/endlsianfv/.
4. Kurgyis, K. *et al.* Uncertainties and robustness with regard to the safety of a repository for high-level radioactive waste: introduction of a research initiative. *Environmental Earth Sciences* **83**, 82, https://doi.org/10.1007/s12665-023-11346-8 (2024).
5. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018, https://doi.org/10.1038/sdata.2016.18 (2016).

6. Ott, F., Elger, K. & Frenzel, S. GFZ Data Services – A domain repository for geosciences data. https://doi.org/10.23689/fidgeo-5570 (2022).
7. Nuclear Energy Agency (NEA). Radioactive Waste Repository Metadata Management (RepMet). https://www.oecd-nea.org/jcms/pl_61001/radioactive-waste-repository-metadata-management-repmet.
8. Nuclear Energy Agency (NEA). Working Party on Information, Data and Knowledge Management (WP-IDKM). https://www.oecd-nea.org/jcms/pl_25233/working-party-on-information-data-and-knowledge-management-wp-idkm.
9. European Joint Programme on Radioactive Waste Management, Brussels (Belgium). EURAD Domain Insight (7.2.2) - Information, Data and Knowledge Management. https://inis.iaea.org/records/7egzr-jhn16.
10. Miron, G. D., Debayle, C., Kudriashova, Y., Hansen, J. & Maugis, V. EURAD-2 - Deliverable 1.3 - Data management plan. https://doi.org/10.5281/zenodo.18174536 (2025).
11. Landesamt für Geologie, Rohstoffe und Bergbau Baden-Württemberg. LGRB Kartenviewer. https://maps.lgrb-bw.de/.
12. Bundesanstalt für Geowissenschaften und Rohstoffe. Borehole map germany. https://boreholemap.bgr.de/mapapps/resources/apps/boreholemap/index.html?lang=en.
13. FIS Geophysik. Geophysics information system (fis-geophysik). https://www.fis-geophysik.de.
14. Weitkamp, A., Frenzel, B., Krumbholz, M., Mrugalla, S. & Sönnke, J. Christa-II - Beschreibung des generischen geologischen Modells für das Endlagersystem "überlagernder ewG". Tech. Rep. 9Y2017020000, Bundesanstalt für Geowissenschaften und Rohstoffe. https://www.bgr.bund.de/DE/Themen/Endlagerung/Downloads/Lanzeitsicherheit/1_Sicherheits_Nachweiskonzepte/2021_12_21_christa_II_ueberlagernder_ewg_ergebnisbericht.pdf?__blob=publicationFile&v=1 (2021).
15. Jahn, S. & Sönnke, J. Endlagerstandmodell Nord (AnSichT) - Teil ii: Zusammenstellung von Gesteinseigenschaften für den Langzeitsicherheitsnachweis. Tech. Rep. 9Y3207000000, Bundesanstalt für Geowissenschaften und Rohstoffe. https://www.bgr.bund.de/DE/Themen/Endlagerung/Downloads/Endlagerforschung/1_Tongestein/2013-01-02_AnSichT_Nord_II.pdf?__blob=publicationFile&v=3 73 pp. (2013).
16. Jahn, S., Mrugalla, S. & Stark, L. Endlagerstandortmodell Süd (AnSichT) - Teil ii: Zusammenstellung von Gesteinseigenschaften für den Langzeitsicherheitsnachweis. Tech. Rep. 9Y3207000000, Bundesanstalt für Geowissenschaften und Rohstoffe. https://www.bgr.bund.de/DE/Themen/Endlagerung/Downloads/Endlagerforschung/1_Tongestein/2016-01-08_AnSichT_Sued_II.pdf?__blob=publicationFile&v=2 (2016).
17. Liu, W., Völkner, E., Minkley, W. & Popp, T. Zusammenstellung der Materialparameter für THM-Modellberechnungen - Ergebnisse aus dem Vorhaben KOSINA. Tech. Rep. 9Y2015010000, Bundesanstalt für Geowissenschaften und Rohstoffe. https://www.bgr.bund.de/DE/Themen/Endlagerung/Downloads/Endlagerforschung/2_Steinsalz/2017-12-01_kosina_materialparameter_fuer_thm_modellberechnungen.pdf?__blob=publicationFile&v=2 (2017).
18. Müller, B., Pippig, U. & Sebastian, U.Geotechnische Klassifikationen von Festgesteinen und Festgebirgen. Springer Spektrum, Berlin, Heidelberg, 1 edn, https://doi.org/10.1007/978-3-662-59777-4 (2019).
19. Minkley, W. *et al*. Beweissicherungsprogramm zum geomechanischen Verhalten von Salinarbarrieren nach starker dynamischer Beanspruchung und Entwicklung einer Dimensionierungsrichtlinie zum dauerhaften Einschluss. Institut für Gebirgsmechanik, *Leipzig*, https://doi.org/10.2314/GBV:63302001X (2010).
20. Prinz, H. & Strauß, R. Ingenieurgeologie Springer Spektrum, Berlin, Heidelberg, 6 edn, https://doi.org/10.1007/978-3-662-54710-6 (2018).
21. Schön, J. H. Physical Properties of Rocks. In Cubitt, J. (ed.) *Handbook of Petroleum Exploration and Production*, vol. 8, i, https://doi.org/10.1016/S1567-8032(11)08012-8 (Elsevier) (2011).
22. Stork, G. Planfeststellungsverfahren zur Stilllegung des Endlagers für radioaktive Abfälle Morsleben. *Procedural document P078, Bundesamt für Strahlenschutz* https://archiv.bge.de/archiv/www.endlager-morsleben.de/Morsleben/DE/mediathek/fachunterlagen/unterlagen-pfv/unterlagen-pfv_node.html (1998).
23. Reinhold, K., Müller, C. & Riesenberg, C. Informationssystem Speichergesteine für den Standort Deutschland - Synthese -. Final report, Bundesanstalt für Geowissenschaften und Rohstoffe, Berlin/Hannover. https://www.bgr.bund.de/DE/Themen/Nutzung-tieferer-Untergrund/Downloads/Speicherkataster_synthese.pdf?__blob=publicationFile&v=1 (2011).
24. Popp, T., Kansy, A., Meyn, R., Pusch, G. & Werunsky, F. Modellentwicklung zur Gaspermeation aus unterirdischen Hohlräumen im Salzgebirge. *Final report, Institut für Erdöl- und Erdgastechnik, Technische Universität Clausthal, Clausthal-Zellerfeld* https://edocs.tib.eu/files/e01fb09/608153567l.pdf (2008).
25. Stober, I. & Jodocy, M. Eigenschaften geothermischer Nutzhorizonte im baden-württembergischen und französischen Teil des Oberrheingrabens. *Grundwasser* **14**, 127–137, https://doi.org/10.1007/s00767-009-0103-3 (2009).
26. Keusen, H., Ganguin, J., Schuler, P. & Buletti, M. Grimesel Test site - Geology. Technical report NTB 87-14E, Nationale Genossenschaft für die Lagerung radioaktiver Abfälle, Zollikofen/Bern https://nagra.ch/wp-content/uploads/2022/08/e_ntb87-014.pdf (1989).
27. Appel, D. & Habler, W. Quantifizierung der Wasserdurchlässigkeit von Gesteinen als Voraussetzung für die Entwicklung von Kriterien zur Grundwasserbewegung - Phase 2: Auswertung der Datensätze für die Kriterienentwicklung. *Tech. Rep., Arbeitskreis Auswahlverfahren Endlagerstandorte* https://dserver.bundestag.de/btd/18/CD09100/4.%20Materialien/K-MAT%2012-02.pdf (2002).
28. Nipp, H.-K. & Heusermann, S. Gebirgsmechanische Beurteilung der Integrität der Salzbarriere im Erkundungsbereich EB1 für das technische Endlagerkonzept 1 (Bohrlochlagerung BSK3). *Results report 9G4141800000, Bundesanstalt für Geowissenschaften und Rohstoffe, Hannover* (2000).
29. Bayer, U., Scheck, M. & Koehler, M. Modeling of the 3D thermal field in the northeast German basin. *Geologische Rundschau* **86**, 241–251, https://doi.org/10.1007/s005310050137 (1997).
30. Kock, I. *et al*. Vorläufige Sicherheitsanalyse für den Standort Gorleben: Integritätsanalyse der geologischen Barriere. *Tech. Rep., Gesellschaft für Anlagen- und Reaktorsicherheit gGmbH, Köln* https://www.bgr.bund.de/DE/Themen/Endlagerung/Downloads/Lanzeitsicherheit/2_Langzeitprognosen_Szenarienanalysen/2012-07-00_Intergritaetsnachweis_geologische_Barriere_VSG_GRS_286.pdf?__blob=publicationFile&v=2 (2012).
31. Lerch, C., Bollingerfehr, W., Filbert, W & Zhang, Q. Thermo-mechanical analyses for evaluating a HLW-repository concept. In Bérest, P., Ghoreychi, M., Hadj-Hassen, F. & Tijani, M. (eds.) *Mechanical Behaviour of Salt VII*, 441–451, https://www.taylorfrancis.com/chapters/edit/10.1201/b12041-60/thermo-mechanical-analyses-evaluating-hlw-repository-concept-lerch-bollingerfehr-filbert-zhang (CRC Press, 2012).
32. Cermak, V. & Rybach, L. 4.1.1 Introductory remarks. In Angenheister, G. (ed.) *Landolt-Börnstein - Group V Geophysics (Subvolume A)*, 305-310, https://doi.org/10.1007/10201894_62 (Springer-Verlag Berlin Heidelberg, 1982).
33. Rath, V. & Clauser, C. Erkennen und Quantifizieren von Strömung: Eine geothermische Rasteranalyse zur Klassifizierung des tiefen Untergrundes in Deutschland hinsichtlich seiner Eignung zur Endlagerung radioaktiver Stoffe - Fortsetzung. *Tech. Rep., Bundesamt für Strahlenschutz*. https://doi.org/10.13140/RG.2.2.36673.99689 (2005).
34. Clauser, C. *et al*. Erkennen und Quantifizieren von Strömung: Eine geothermische Rasteranalyse zur Klassifizierung des tiefen Untergrundes in Deutschland hinsichtlich seiner Eignung zur Endlagerung radioaktiver Stoffe. *Final report, Bundesamtes für Strahlenschutz*. https://doi.org/10.13140/RG.2.2.36673.99689 (2002).
35. Kneuker, T. & Furche, M. Capturing the structural and compositional variability of Opalinus Clay: constraints from multidisciplinary investigations of Mont Terri drill cores (Switzerland). *Environmental Earth Sciences* **80**, 421, https://doi.org/10.1007/s12665-021-09708-1 (2021).

36. Gélis, C. *et al.* Ability of High-Resolution Resistivity Tomography to Detect Fault and Fracture Zones: Application to the Tournemire Experimental Platform, France. *Pure and Applied Geophysics* **173**, 573–589, https://doi.org/10.1007/s00024-015-1110-1 (2016).
37. Reinhold, K., Stark, L., Kühnlenz, T. & Ptock, L. Endlagerstandmodell Süd (AnSichT) - Teil i: Beschreibung des Endlagerstandortmodells. *Tech. Rep*. 9Y3207000000, Bundesanstalt für Geowissenschaften und Rohstoffe https://www.bgr.bund.de/DE/Themen/Endlagerung/Downloads/Endlagerforschung/1_Tongestein/2016-07-26_AnSichT_Sued_I.pdf?__blob=publicationFile&v=2 (2016).
38. NAGRA. Sondierbohrung Benken Untersuchungsbericht. Technical report NTB 00-01, Nationale Genossenschaft für die Lagerung radioaktiver Abfälle https://nagra.ch/wp-content/uploads/2022/08/d_ntb00-001-Textband.pdf (2001).
39. Bossart, P. & Thury, M. Mont Terri Rock Laboratory Project, Programme 1996 to 2007 and Results. Report No. 3, Swiss Geological Survey https://www.mont-terri.ch/dam/fr/sd-web/ClxkQzU7GNhg/BLGD-03.pdf (2008).
40. Thury, M. & Bossart, P. Mont Terri Rock Laboratory: Results of the Hydrogeological, Geochemical and Geotechnical Experiments Performed in 1996 and 1997. Report No. 23, Swiss National Hydrological and Geological Survey https://www.mont-terri.ch/dam/fr/sd-web/06eNs9TWwE7-/GBD-23.pdf (1999).
41. Jodocy, M. & Stober, I. Porositäten und Permeabiliotäten im Oberrheingraben und Südwestdeutschen Molassebecken **127**, 20–27 (2011).
42. Marschall, P. *et al.* Synthesis of Hydrogeological Investigations at the Mont Terri site (Phases 1 to 5) In Heitzmann, P. (ed.) *Mont Terri Project – Hydrogeological Synthesis, Osmotic Flow. Geology Series No. 6 (Federal Office for Water and Geology (FOWG))* https://www.mont-terri.ch/dam/fr/sd-web/DumoFAfwKxPx/BWGD-6.pdf (2004).
43. Reinhold, K., Jahn, S., Kühnlenz, T., Ptock, L. & Sönnke, J. Endlagerstandmodell Nord (AnSichT) - Teil i: Beschreibung des geologischen Endlagerstandmodells. Tech. Rep. 9Y3207000000, Bundesanstalt für Geowissenschaften und Rohstoffe https://www.bgr.bund.de/DE/Themen/Endlagerung/Downloads/Endlagerforschung/1_Tongestein/2013-01-02_AnSichT_Nord_I.pdf?__blob=publicationFile&v=3 72 pp. (2013).
44. Jobmann, M. *et al.* Untersuchungen zur sicherheitstechnischen Auslegung eines generischen Endlagers im Tonstein in Deutschland -GENESIS- Abschlussbericht. *Final report, DBE Technology GmbH* https://www.bge-technology.de/fileadmin/user_upload/FuE_Berichte/Genesis/GENESIS-_Generisches_Endlager_im_Tonstein.pdf.pdf (2007).
45. Schnier, H. Ergebnisse der festigkeitsmechanischen Laboruntersuchungen an Gesteinsproben aus dem Hangend- und Liegend-Bereich der Grube Konrad. *Laboratory report 11914/86, Bundesanstalt für Geowissenschaften und Rohstoffe, Hannover* (1986).
46. Schnier, H. Ergebnisse festigkeitsmechanischer Laborsuchungen an Gesteinsproben aus den Schachtwiderlagerbohrungen im Schacht Konrad 1. *Laboratory report 10739/86, Bundesanstalt für Geowissenschaften und Rohstoffe, Hannover* (1987).
47. Schnier, H. Ergebnisse der festigkeitsmechanischen Laboruntersuchungen an Proben aus den Schachtwiderlagerbohrungen im Schacht Konrad 2. *Laboratory report 10739/86, Bundesanstalt für Geowissenschaften und Rohstoffe, Hannover* (1986).
48. Müller, K., Kreutz, H., Ziegenmeyer, H. & Meier, G.-A. Probenuntersuchungen an Lager-Begrenzungsbohrungen der Schachtanlage Konrad. *Abschlussbericht EU 120, Clausthal University of Technology, Clausthal-Zellerfeld* https://www.bge.de/fileadmin/user_upload/Konrad/Wesentliche_Unterlagen/Geologie_Hydrogeologie/EU_120_geschwaerzt_nicht_barrierefrei.pdf (1987).
49. BfS. Planfeststellungsbeschluss für die Errichtung und den Betrieb des Bergwerkes Konrad in Salzgitter als Anlage zur Endlagerung fester oder verfestigter radioaktiver Abfälle mit vernachlässigbarer Wärmeentwicklung vom 22. Mai 2002. *Planfeststellungsbeschluss, Bundesamt für Strahlenschutz, Salzgitter* https://www.bge.de/fileadmin/user_upload/Konrad/Wesentliche_Unterlagen/Genehmigungsunterlagen/Planfeststellungsbeschluss_Endlager_Konrad_vom_22_Mai_2002.pdf (2002).
50. Hoth, P. *et al.* Endlagerung radioaktiver Abfälle in tiefen geologischen Formationen Deutschlands: Untersuchung und Bewertung von Tongesteinsformationen. *Tech. Rep., Bundesanstalt für Geowissenschaften und Rohstoffe* https://www.bgr.bund.de/DE/Themen/Endlagerung/Downloads/Endlagerforschung/1_Tongestein/2007-04-00_BGR_Tonstudie2007.pdf?__blob=publicationFile&v=1 (2007).
51. Yaramanci, U. Geoelectric exploration and monitoring in rock salt for the safety assessment of underground waste disposal sites. *Journal of Applied Geophysics* **44**, 181–196, https://doi.org/10.1016/S0926-9851(99)00013-0 (2000).
52. Yaramanci, U. Relation of in situ resistivity to water content in salt rocks. *Geophysical Prospecting* **42**, 229–239, https://doi.org/10.1111/j.1365-2478.1994.tb00207.x (1994).
53. LGRB-Informationen 20 (Bearbeitung: Plum, H., Ondreka, J. & Armbruster, V.). Hydrogeologische Einheiten in Baden-Württemberg. Tech. Rep. Landesamt für Geologie, Rohstoffe und Bergbau Baden-Württemberg https://produkte.lgrb-bw.de/docPool/c624_data.pdf (2008).
54. Bollingerfehr, W. *et al.* KOSINA Concept developments for a generic repository for heat-generating waste in bedded salt formations in Germany. *Synthesis Report, BGE TECHNOLOGY GmbH* https://www.bgr.bund.de/DE/Themen/Endlagerung/Downloads/Endlagerforschung/2_Steinsalz/2018-10-30_kosina_concept_developments_bge_tec.pdf?__blob=publicationFile&v=1 (2018).
55. Völkner, E., Kühnlenz, T., Hammer, J. & Gast, S. Entwicklung generischer geologischer Modelle für flach lagernde Salzformationen — Ergebnisse aus dem Vorhaben KOSINA. *Results Report 9Y2015010000, Bundesanstalt für Geowissenschaften und Rohstoffe* https://www.bgr.bund.de/DE/Themen/Endlagerung/Downloads/Endlagerforschung/2_Steinsalz/2017-10-12_Kosina_EB_Generische_geologische_Modelle.pdf?__blob=publicationFile&v=2 (2017).
56. Falcon-Suarez, I. H. & Dale, M. Geophysical and transport properties of salt rocks subjected to loading/unloading, and during dissolution. *NERC EDS National Geoscience Data Centre* https://doi.org/10.5285/13c05fe8-0d1d-49e5-b55e-5cf4ef241c70 Dataset (2023).
57. Alheid, H.-J. Seismische Lastannahme für unter Tage, Ableitung der Bewegungsgrößen unter Tage. *Tech. Rep. 99586, Bundesanstalt für Geowissenschaften und Rohstoffe, Hannover* (1986).
58. Trautwein, U. Poroelastische Verformung und petrophysikalische Eigenschaften von Rotliegend Sandsteinen https://depositonce.tu-berlin.de/items/urn:nbn:de:kobv:83-opus-11515 (2005).
59. Huysmans, M. & Dassargues, A. Hydrogeological modeling of radionuclide transport in low permeability media: a comparison between Boom Clay and Ypresian Clay. *Environmental Geology* **50**, 122–131, https://doi.org/10.1007/s00254-006-0191-7 (2006).
60. Bonetto, S. *et al.* Study of the Mechanical Properties of a Conglomerate. *Procedia Engineering* **158**, 248–253, https://doi.org/10.1016/j.proeng.2016.08.437 (2016).
61. Müller-Huber, E. & Börner, F. Multi-parameter petrophysical characterization of Variscan greisen rocks from the Southern Bohemian Batholith (Austria) and the Eastern Erzgebirge Volcano-Plutonic Complex (Germany). *Austrian Journal of Earth Sciences* **110**, https://doi.org/10.17738/ajes.2017.0006 (2017).
62. Heap, M. J. *et al.* Petrophysical properties of the Muschelkalk from the Soultz-sous-Forêts geothermal site (France), an important lithostratigraphic unit for geothermal exploitation in the Upper Rhine Graben. *Geothermal Energy* **7**, 27, https://doi.org/10.1186/s40517-019-0145-4 (2019).
63. Menzel, N., Klitzsch, N., Altenbockum, M., Müller, L. & Wagner, F. M. Prospection of faults on the Southern Erftscholle (Germany) with individually and jointly inverted refraction seismics and electrical resistivity tomography. *Journal of Applied Geophysics* **231**, 105549, https://doi.org/10.1016/j.jappgeo.2024.105549 (2024).
64. Behrens, C. *et al.* TransPyREnd: A code for modelling the transport of radionuclides on geological timescales. In *Advances in Geosciences* **58**, 109–119, https://doi.org/10.5194/adgeo-58-109-2023 (2023).

65. Chen, Q., Boxberg, M. S., Menzel, N., Wagner, F. M. & Kowalski, J. Radioactive Contamination Risk Assessment in Long-Term Radioactive Waste Disposal: Actionable Data-Hub for Analysis-Readiness in Process and Impact Models, EGU General Assembly 2025, Vienna, Austria, 27 Apr–2 May 2025, EGU25-18973, https://doi.org/10.5194/egusphere-egu25-18973 (2025).
66. Menzel, N., Uhlemann, S. & Wagner, F. M. Optimized experimental design strategies for ERT monitoring of transient flow processes. *Geophysical Journal International* **244**, ggaf523, https://doi.org/10.1093/gji/ggaf523 (2026).
67. YAML Language Development Team. Yaml ain't markup language (yaml)— version 1.2, revision 1.2.2, https://yaml.org/spec/1.2.2/.
68. pandas development team, T. pandas-dev/pandas: Pandas (2025).
69. International Atomic Energy Agency. Geological disposal of radioactive waste: Technological implications for retrievability. *Tech. Rep., International Atomic Energy Agency* https://www.iaea.org/publications/8022/geological-disposal-of-radioactive-waste-technological-implications-for-retrievability. **64** pp. (2009).
70. de la Varga, M., Schaaf, A. & Wellmann, F. GemPy 1.0: open-source stochastic geological modeling and inversion. *Geoscientific Model Development* **12**, 1–32, https://doi.org/10.5194/gmd-12-1-2019 (2019).
71. Sullivan, C. B. & Kaszynski, A. A. PyVista: 3D plotting and mesh analysis through a streamlined interface for the Visualization Toolkit (VTK). *Journal of Open Source Software* **4**, 1450, https://doi.org/10.21105/joss.01450 (2019).
72. Commission for the Management of Geoscience Information (CGI). Cgi vocprez — simple lithology classifier (geosciml). Vocabulary / GeoSciML. https://cgi.vocabs.ga.gov.au/object?uri=http%3A//resource.geosciml.org/classifiersc580heme/cgi/2016.01/simplelithology 2025-08-31.
73. British Geological Survey. Simple lithology. https://www.bgs.ac.uk/technologies/bgs-rock-classification-scheme/.
74. McKie, T. Paleogeographic Evolution of Latest Permian and Triassic Salt Basins in Northwest Europe. In Juan, I. S., Joan, F. F. & Gabor, T. (eds.) *Permo-Triassic Salt Provinces of Europe, North Africa and the Atlantic Margins*, 159–173, https://doi.org/10.1016/B978-0-12-809417-4.00008-2 (Elsevier, 2017).
75. Bundesgesellschaft für Endlagerung mbH. Zwischenbericht Teilgebiete gemäß 13 StandAG. Tech. Rep. SG01101/16-1/2-2019#3, Bundesgesellschaft für Endlagerung mbH Objekt-ID: 755925 - Revision: 000 (2020).
76. Bachmann, G. *et al*. Depositional history and sedimentary cycles in the Central European Basin System. In Littke, R., Bayer, U., Gajewski, D. & Nelskamp, S. (eds.) *Dynamics of Complex Intracontinental Basins: The Example of the Central European Basin System*, 156–172 (Springer, Potsdam, 2008).
77. Maystrenko, Y., Bayer, U., Brink, H.-J. & Littke, R. The Central European Basin System - an Overview. In Littke, R., Bayer, U., Gajewski, D. & Nelskamp, S. (eds.) *Dynamics of Complex Intracontinental Basins: The Central European Basin System*, 16–34, https://doi.org/10.1007/978-3-540-85085-4_2 (Springer, Berlin, Heidelberg 2008).
78. Ahlrichs, N. *et al*. Impact of Late Cretaceous inversion and Cenozoic extension on salt structure growth in the Baltic sector of the North German Basin. *Basin Research* **34**, 220–250, https://doi.org/10.1111/bre.12617 (2022).
79. Caglayan, D. G. *et al*. Technical potential of salt caverns for hydrogen storage in Europe. *International Journal of Hydrogen Energy* **45**, 6793–6805, https://doi.org/10.1016/j.ijhydene.2019.12.161 (2020).
80. Minougou, J. D., Gholami, R. & Andersen, P. Underground hydrogen storage in caverns: Challenges of impure salt structures. *Earth-Science Reviews* **247**, 104599, https://doi.org/10.1016/j.earscirev.2023.104599 (2023).
81. Soto, J. I., Flinch, J. F. & Tari, G. C. Permo-Triassic Salt Provinces of Europe, North Africa and the Atlantic Margins: Tectonics and Hydrocarbon Potential, https://doi.org/10.1016/C2015-0-05796-3 (Elsevier, 2017)
82. Van Gent, H., Urai, J. L. & De Keijzer, M. The internal geometry of salt structures - A first look using 3D seismic data from the Zechstein of the Netherlands. *Journal of Structural Geology* **33**, 292–311, https://doi.org/10.1016/j.jsg.2010.07.005 (2011).
83. Menaceur, H., Delage, P., Tang, A.-M. & Conil, N. The thermo-mechanical behaviour of the Callovo-Oxfordian claystone. *International Journal of Rock Mechanics and Mining Sciences* **78**, 290–303, https://doi.org/10.1016/j.ijrmms.2015.07.002 (2015).
84. Bonin, B. Deep geological disposal in argillaceous formations: studies at the Tournemire test site. *Journal of Contaminant Hydrology* **35**, 315–330, https://doi.org/10.1016/S0169-7722(98)00132-6 (1998).
85. Armand, G., Conil, N., Talandier, J. & Seyedi, D. M. Fundamental aspects of the hydromechanical behaviour of Callovo-Oxfordian claystone: From experimental studies to model calibration and validation. *Computers and Geotechnics* **85**, 277–286, https://doi.org/10.1016/j.compgeo.2016.06.003 (2017).
86. Meschede, M. Deutschland im Perm und Mesozoikum. In Meschede, M. (ed.) *Geologie Deutschlands: Ein prozessorientierter Ansatz*, 103–159, https://doi.org/10.1007/978-3-662-56422-6_11 (Springer, Berlin, Heidelberg 2018).
87. Jordan, P. *et al*. The middle to late Triassic Bänkerjoch and Klettgau formations of northern Switzerland. *Swiss Journal of Geosciences* **109**, 257–284, https://doi.org/10.1007/s00015-016-0218-3 (2016).
88. Hassine, S. B. H., Davies, C. & Garbil, R. Forty-five years of joint research programmes on geological disposal of radioactive waste and the pioneering role of the HADES Underground Research Laboratory. *Geological Society, London, Special Publications* **536**, 225–236, https://doi.org/10.1144/SP536-2021-205 (2023).
89. Li, X., Neerdael, B., Raymaekers, D. & Sillen, X. The construction of the HADES underground research laboratory and its role in the development of the Belgian concept of a deep geological repository. *Geological Society, London, Special Publications* **536**, 159–184, https://doi.org/10.1144/SP536-2022-101 (2023).
90. Kopp, J. & Bankwitz, P. Die Europäische Kristallinzone (EKZ) - eine Übersicht. *Z. geol. Wiss.* **31**, 179–196 (2003).
91. Nasir, S., Okrusch, M., Kreuzer, H., Lenzt, H. & Höhndorf, A. Geochronology of the Spessart Crystalline Complex, Mid-German Crystalline Rise. *Mineralogy and Petrology* **44**, 39–55, https://doi.org/10.1007/BF01167099 (1991).
92. Souček, K. *et al*. Geotechnical Characterization of Bukov Underground Research Facility. *Procedia Engineering* **191**, 711–718, https://doi.org/10.1016/j.proeng.2017.05.236 (2017).
93. Aminzadeh, A., Petruzalek, M., Vavrycuk, V. & Lokajicek, T. The pressure dependent elastic anisotropy of Grimsel granite and Bukov paragneiss. *IOP Conference Series: Earth and Environmental Science* **1124**, 012049, https://doi.org/10.1088/1755-1315/1124/1/012049 (2023).
94. Petruzalek, M., Aminzadeh, A., Lokajicek, T. & Vavrycuk, V. Elastic Anisotropy of Typical Rocks at Grimsel and Bukov URLs. In *57th U.S. Rock Mechanics/Geomechanics Symposium, ARMA–2023–0415 ARMA, Atlanta, Georgia, USA*, https://onepetro.org/ARMAUSRMS/proceedings/ARMA23/ARMA23/ARMA-2023-0415/532325 (2023).
95. Davis, K. R., Peabody, B. & Leach, P. Universally Unique IDentifiers (UUIDs). Request for Comments RFC 9562, Internet Engineering Task Force. https://doi.org/10.17487/RFC9562 (2024).
96. Chen, Q., Menzel, N., Boxberg, M. S. & Kowalski, J. Dataset for the smart data hub: Subsurface properties, https://doi.org/10.5281/zenodo.19886769 (2026).
97. Chen, Q., Menzel, N., Boxberg, M. S. & Kowalski, J. Smart data hub: Integrating subsurface properties into model-based decision-support tools, https://doi.org/10.5281/zenodo.19922612 (2026).
98. Clark, C. E. Letter to the Editor—The PERT Model for the Distribution of an Activity Time. *Operations Research* **10**, 405–406, https://doi.org/10.1287/opre.10.3.405 (1962).
99. Menning, M. & Hendrich, A. Stratigraphic Table of Germany Compact 2022, https://www.stratigraphie.de/ergebnisse/STGCurldate=2026-04-25 (STGC 2022).
100. Cohen, K. M., Finney, S. C., Gibbard, P. L. & Fan, J.-X. The ICS International Chronostratigraphic Chart. *Episodes Journal of International Geoscience* **36**, 199–204, https://doi.org/10.18814/epiiugs/2013/v36i3/002 (2013).
101. Plotly Technologies Inc. Collaborative data science https://plot.ly (2015).

Acknowledgements
We would like to thank Jan von Harten from RWTH Aachen University for his assistance with the GemPy software.

Author contributions
Q.C.: Conceptualization, Methodology, Software, Validation, Data curation, Formal analysis, Writing - Original Draft, Writing - Review & Editing, Visualization. N.M.: Data curation, Writing - Original Draft, Writing - Review & Editing. M.B.: Conceptualization, Methodology, Writing - Review & Editing, Supervision. J.K.: Conceptualization, Methodology, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

Funding
We acknowledge the contribution of RWTH Aachen University to the Open Access funding within the framework of the DEAL initiative. This work was supported through the Uncertainties and Robustness with regard to the Safety of a Repository for High-Level Radioactive Waste (URS) research cluster, which was funded by the Federal Company for Radioactive Waste Disposal (BGE) (Grant No.: STAFuE-21-4-Klei). Open Access funding enabled and organized by Projekt DEAL.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Q.C. or J.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2026