Sub
Sol

Draft Data Management Plan

SUBSOL has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 642228

Title:	Draft Data Management Plan

Grant agreement no:	642228
Work Package:	WP5
Deliverable number:	D51
Partner responsible:	KWR
Deliverable author(s):	Gerard van Den Berg (KWR), Christos Makropoulos (KWR, NTUA), George Karavokiros (NTUA)
Quality assurance:	Christos Makropoulos (KWR, NTUA)
Planned delivery date:	18 April 2017
Actual delivery date:	18 April 2017
Dissemination level:	PU *PU = Public* *PP = Restricted to other programme participants (including the Commission Services)* *RE = Restricted to a group specified by the consortium (including the Commission Services)* *CO = Confidential, only for members of the consortium (including the Commission Services)*

Table of contents

Executive Summary

The purpose of this Data Management Plan (DMP) is to provide an analysis of the main elements of the data management policy that will be used by the SUBSOL project with regard to all the datasets that will be generated or collected by the project. The DMP is not a fixed document, but evolves during the lifespan of the project; in fact it functions as a dynamic document of agreements.

SUBSOL is aiming to provide robust, effective, sustainable, and cost-efficient answers to freshwater resources challenges in coastal areas worldwide through a market breakthrough of SWS. In order to accomplish that, it is necessary to develop practical approaches which will accelerate acceptance of SWS and will broaden the market reach and uptake. These approaches consist of continuation and replication of successful full scale pilots, development of decision support tools and business cases, assessment of market readiness and of policy and legal framework conditions, and capacity building activities. As such the project will collect and generate (broadly) two types of data: (a) data related to the pilots themselves and (b) data related to SWS design, implementation and uptake. The DMP explains the differences between these two types of data and explains how the project intends to manage them. It presents the project’s strategy on standardisation, archiving and data sharing and provides an overview of all key aspects related to data within the SUBSOL project.

1. Introduction

1.1 What is the Data Management Plan

The purpose of the Data Management Plan (DMP) is to provide an analysis of the main elements of the data management policy that will be used by the SUBSOL project with regards to all the datasets that will be generated by the project. The DMP is not a fixed document, but evolves during the lifespan of the project; in fact it functions as a dynamic document of agreements. The DMP should address the points presented below on a dataset by dataset basis and should reflect the current status of reflection within the consortium about the data that will be generated, collected, stored and processed.

In principle, publicly funded research data are a public good, produced for the public interest that should be made openly available with as few restrictions as possible in a timely and responsible manner that does not harm intellectual property. On this basis, the DMP intends to help researchers consider at an early stage, when research is being designed and planned, how data will be managed during the research process and shared afterwards with the wider research community.

The benefits of a well-designed DMP not only concern the way data are treated but also the successful outcome of the project itself. A properly planned DMP guides the researchers first to think what to do with the data and then how to collect, store and process them, etc. Furthermore, a planning in data treatment is important for addressing timely security, privacy and ethical aspects. This way the research data are kept in track in cases of possible staff or other changes. The DMP can also increase preparedness for possible data requests. In short, planned activities, such as implementation of well-designed DMP, stand a better chance of meeting their goals than unplanned ones.

The process of planning is also a process of communication, increasingly important in a multi-partner research. The characteristics of collaboration should be accordingly harmonised among project partners from different organisations or different countries. The DMP also provides an ideal opportunity to engender best practice with regards to e.g. file formats, metadata standards, storage and risk management practices, leading to greater longevity and sustainability of data and higher quality standards.

Ultimately, the DMP should engage researchers in conversations with those providing the services. In this context, the DMP becomes a document in accordance with relevant standards and community best practice. Data should be shared, edited, and monitored among those contributing to the project. Releasing research data should follow legal, ethical and commercial terms and conditions. To serve the multiple purposes just described, the DMP should be designed for easy digital exchange across a variety of applications. The best way to approach this in today's complex world of information technology is through a metadata standard describing a data model of elements constituting the DMP.

1.2 Relevant EU policies

From 2014, EU is working towards establishing an open access policy for research data. In order to accomplish that, it calls for a Data Management Plan report from Horizon 2020 participants. The DMP are required for 'key areas' of the Horizon 2020 programme, covered by the Open Data Pilot. These include several technology-oriented strands and others addressing 'societal challenges'. Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020[1] echo the G8 Science Ministers' statement (2013)[2], which offered similar good practice principles.

According to EU guidelines, the DMP should address a number of key aspects of data management within a project. In particular, it should state whether the data and associated software produced and/or used in the project are identifiable by means of a standard identification. The accessibility of software and data in use during the implementation of the project (e.g. licensing framework for research and education, embargo periods, commercial exploitation, etc.), the usability of the stored data by third parties, the data exchange between researchers, institutions, organisations, countries, etc., as well as the description of quality requirements, should be described in the DMP according to the EU policies. *SUBSOL expects to participate in the EU initiative of the Open Research Data Pilot* (see below).

Since the DMP is expected to mature during the project, more developed versions of the plan will be produced at later stages.

1.3 The Open Research Data Pilot

Open data is data that is free to access, reuse, repurpose, and redistribute. The Open Research Data Pilot aims to make the research data generated by selected Horizon 2020 projects accessible with as few restrictions as possible, while at the same time protecting sensitive data from inappropriate access. SUBSOL is participating in the EU Open Research Data Pilot.

The types of data concerned are:

- Data (including associated metadata) needed to validate the results presented in scientific publications ("underlying data")
- Data collected from sensors in SUBSOL deployments
- Data concerning Intellectual Property Rights (IPRs)
- Other data (including associated metadata) as specified in this data management plan

Regarding the digital research data generated in the action the beneficiaries participating in the pilot will:

- Deposit data in a research data repository; that is the project's Knowledge Base (KB), (Task 3.1)

[1] Guidelines on Open Access to Scientific Publications & Research Data in Horizon2020, ver.1.0, p. 8-11,11 Dec 2013

[2] G8 Science Ministers Statement, Foreign & Commonwealth Office, 13 June 2013, UK

- Provide information, via the KB and the project Toolkit (Task 3.2), about tools and instruments at the disposal of the beneficiaries, also necessary for validating results obtained (where possible, provide the tools and models themselves).
- Take measures to make it possible for third parties to access, mine, exploit, reproduce and disseminate free of charge for all users (using e.g. Creative Commons Licenses)

2 Project Context

2.1 The SUBSOL project in brief

SUBSOL is aiming to provide robust, effective, sustainable, and cost-efficient answers to freshwater resources challenges in coastal areas worldwide through a market breakthrough of SWS. In order to accomplish that, it is necessary to develop practical approaches which will accelerate acceptance of SWS and will broaden the market reach and uptake. These approaches consist of continuation and replication of successful full scale pilots, development of decision support tools and business cases, assessment of market readiness and of policy and legal framework conditions, and capacity building activities.

The well-monitored reference sites (WP1) form an excellent basis for building long-term SWS track records and for addressing questions related to scaling-up SWS. Thus, existing data of operational sites will be collected and included in the SWS Knowledge Base (KB) that will be developed in the scope of the project.

To demonstrate the robustness of SWS and their applicability under different environmental and societal conditions (WP2), four field pilots in Denmark, Greece, the Netherlands and the USA will be implemented. As these replications will be tailor-made to the local conditions, serving the needs of the local clients, decision makers and other stakeholders, information that will be generated in the scope of the pilots will also be stored and processed in the project SWS Knowledge Base (KB) in an explicit and well-structured manner. This web-based knowledge repository will be developed within task 3.1 and will provide information on what is available, what has already been applied and tested under different contexts, what were the main lessons learned and what are the key legislative and regulatory issues that need to be considered. In particular, different SWS experiences and practices, technical and regulatory contexts, as well as policy analysis and policy recommendations will be recorded. The material that complements and supports SWS implementations includes reports, peer reviewed journals, best practice guidelines, datasets, audio-visual material. Furthermore, important information on economic viability of the SWS implementations will be included in the repository.

A dedicated SWS toolkit will also be developed (task 3.2) that will include hydro-technical SWS decision support tools to guide the choice and initial design of appropriate SWS solutions for each specific problem. The tools will also provide data (if available) on key technical figures e.g. initial design specifications, costs or even locations (if GIS data are available) at a prefeasibility level for the preferred options and provide an initial risk assessment.

Materials to develop training packages to help decision makers with making informed decisions about investments and securing future water supply will be collected in task 3.3, including technical guidance for SWS implementations, regulation/legislation, business case and business development advice, experiences and funding and related support.

Regarding the commercialisation and market penetration (task 4.1) a step-by-step list of activities for future use of SME technology providers will be implemented; a market scan to assess the opportunity on national or regional market entry will be developed and applied, through an easily applicable checklist of criteria.

Additionally, a more comprehensive market analysis including a gap analysis of target markets and identification of implementation barriers to assess the specific technology and management viability will be carried out (task 4.2). The analysis will be transferred to SWS stakeholders through policy briefs. Recommendations for stakeholder engagement approach will be produced as “solution packages” for decision-makers that summarise advantages, approach and set-up of proposed technologies, including recommended framework conditions.

To improve the visibility and commercialization of SWS a general strategic approach is developed that includes target group assessment and definition of specific sets of knowledge products for case sites and technology application (task 4.3).

SUBSOL will share experiences and outcomes with stakeholder groups through its online platform that will be linked to existing networks, including EIP Water.

In terms of data availability, the intention of SUBSOL is to produce freely available research data, findings, results, reports and deliverables, and share them with the minimal restrictions, aside from the attribution to the authors or creators. With regards to produced software, programming code, interface components, etc., the IP and copyright of the organisations that produced it will be recorded. However, these organisations will be encouraged to make the software open source via an open software license.

Additional project information, including project events, meetings, new findings, deliverables released etc. will be widely disseminated to all interested parties through the project’s website.

2.2 Roles & responsibilities within SUBSOL

The principle role of the project coordinator is to design and oversee the research implemented; whereas the responsibilities of the research staff include designing research but also collecting, processing and analysing data as well as knowing where to keep the data and who has access to it. They are also responsible for generating metadata and relevant documentation. The database designers will cooperate with the research staff to properly house, archive and share the data collected. Additional external staff will be responsible for data collection and data entry, if needed, as well as transcription, processing, analysis, and standard protocols agreements. The role of administration staff is to manage and administrate research and funding as well as to review ethical issues of the project. Institutional IT services are responsible for data storage, security and backup services, while external data centres can facilitate data sharing if needed.

2.3 The SUBSOL data management approach

The SUBSOL data management approach consists of several key issues that will be discussed in the following sections.

- Data categories
- Existing data/New data
- Standards
- Backup and security of data
- Sharing and archiving data

2.4 Data categories

This section presents the major data categories to be generated or collected in the project: (a) project information, (b) knowledge information, (c) field data and (d) applications source code. Another category referring to the metadata of geospatial information is described in section **Error! Reference source not found.**.

2.5 SUBSOL Website

The partners in the SUBSOL consortium decided early to setup a project-related website. The website will serve the project in various ways, including the following:

- It will be the official reference website of the project, providing information about its mission and the general approach, the project partners and the development status. The URL of the project's homepage is: http://www.subsol.org/

- It will link to all other components of the SUBSOL project (e.g. Knowledge Base, Toolbox, Market Place).

- It will provide a blog area with regularly updated news about the project. Later in the project the developed SUBSOL software components will be announced.

- A dedicated area for downloads will be used to publish reports and white papers. All documents are published using the portable document format (PDF). All downloads are enriched by using simple metadata information like the title and the type of the document.

The website will be backed up on a regular basis. All information on the SUBSOL website will be accessible without restrictions. The information will be indexed by web search engines like Microsoft Bing or Google Search and statistics regarding the website traffic will be generated and analysed frequently.

2.6 Data management in the Knowledge Base

The Knowledge Base (KB) will store a major part of the data produced by the project. The focus will be on the following main categories (entities):

Measures: Measures and solutions for advanced and sustainable water resources management that allows for an enhanced protection and utilization of the freshwater resources in coastal areas. Managed Aquifer Recharge (MAR) technologies and Subsurface Water Solutions (SWS).

Case studies: Applied configurations of MAR and SWS techniques.

Tools: Tools that have been designed for the purposes of MAR/SWS implementation.

Complementing data collected in these main categories, additional data will be stored in the SWS Knowledge Base: **Regulations**, giving the legal framework under which the measures in case studies are consistent with, issued **Publications** that refer to MAR / SWS configurations as well as related case studies, tools and regulations and **Legal entities** (authorities, companies, organisations etc.) that are responsible for implementing measures in specific case studies or own specific tools.

Limited information on **Projects** will be collected that will allow to relate EU and other projects implementing MAR/SWS techniques with the respective case studies and tools stored in the Knowledge Base. **Illustrations** of images or schematic representations of measures or of case study configurations will be collected in formats suitable for publishing over the Web (e.g. png, jpg) as well as associated metadata such as legend or source.

It is planned that information from the above data categories will be entered in the Knowledge Base by authorized users through custom online forms. The data importing system will ensure the validity of the data in real time. Typical checks will be implemented, including data type check, duplicate checks, mandatory and conditional mandatory data checks and range checks.

The Knowledge Base will combine the collected data producing new information. All information will be published over the web in form of tables, lists, graphs, reports etc., provided that no licensing restrictions apply.

A backup and recovery plan of the Knowledge Base will be developed (see also Section 0) and tested ensuring the recovery of the data in case of an emergency situation e.g. system failure or cyber-attack.

2.7 Data Monitoring System

The SUBSOL Data Monitoring System will form part of the toolbox (developed in task 3.2) and it will be a versatile and portable web application that will collect and display (a) *meteorological*, (b) *hydrogeological* and (c) *environmental* data. It will store data in a database and visualize them in the form of a Dashboard. The Dashboard will be developed using open-source technologies and will be modular and fully extensible. A set of interfaces will be developed, supporting integration of data from different formats (e.g. csv, raw data). Data collected per pilot site will be stored in the premises of each pilot, while a set of security and privacy mechanisms are going to be deployed for accessing and exchanging such data.

Stations gather meteorological, hydrological and environmental data. These (raw) data from the stations are, then, transferred through data exchange files. The *Data collector* is responsible for collecting these data files. Unfortunately, in most cases data exchange file formats do not follow a specific standard. For example, some stations may export JSON files, other stations may export CSV files, and in some cases that stations produce XML or simple unstructured text files of some other format. Even namespaces are may be different.

Obviously, there is no way to cover the whole range of file formats and data structures. Therefore, upon receiving a file, the parser has to read it and support the user to associate data from the source (station/sensor) with the data fields in the database. Once the data are in the database, they can be visualized in a variety of ways (time series, charts, tables etc.) through a user friendly GUI.

The SUBSOL Data Monitoring System will be able to support interconnection of collected data with data available in the Linked Open Data (LOD) Cloud as well as with data available as open data in other formats. For this purpose, the workbench made available through the LinDA project (http://linda.epu.ntua.gr/) can be used.

2.8 Version control system

Various information systems will have to be developed during the lifetime of the SUBSOL project. Many of them will require the collaborative development of software applications, websites and other computer programs either from scratch or at least significant parts of them. The management and version control of the source code may become a significant problem, especially in large software development projects involving several developers.

The SUBSOL IT team will systematically use **Git** as the source code management system for software development. Git will allow the collaborate development of the software providing effective distributed version control. It is easy to use and yet powerful and efficient to handle even large projects.

A copy of each Git repository will be uploaded ("pushed") to a publicly available Git server such as GitHub or Bitbucket. They are both well-established online servers which support distributed source code development, management, and revision control. They enable world-wide collaboration between developers and also provide some additional facilities to work on documentation and to track issues. GitHub provides paid and free service plans. Free service plans can have any number of public, open-access repositories with unlimited collaborators. Private, non-public repositories require a paid service plan while Bitbucket allows private repositories to be shared with up to five collaborators. The platforms use metadata like contributors' nicknames, keywords, time, and data file types to structure the projects and their results. The terms of service state that no intellectual property rights are claimed for provided material.

2.9 Publications

The publications to be produced within the project refer to Managed Aquifer Recharge (MAR) technologies and Subsurface Water Solutions (SWS). Publication types include journals, books, project reports, scientific articles, grey literature, webpages, guidelines, tutorials etc. Metadata for this type of product will include the authors and the title of the publication, the publisher, the year, the url, relevant keywords as well as an abstract of the publication.

All data created or provided by organisations or stakeholders will maintain the copyright and intellectual property of the data providers or data creators in compliance with the data providers own terms and conditions.

The project will make use of three different possibilities for open access provision, depending upon what is most appropriate for the publications selected, the article itself and the partners that have produced the material.

1. Publication in open access papers - papers that provide open access immediately by default.

2. Publication in papers whereby authors pay a fee to publish the material as open access immediately. Most high-level journals offer this option.

3. Publication whereby authors archive the material in a disciplinary, institutional or public repository.

2.10 Secondary use of personal data

With regards to the Protection of Personal Data raised in the SUBSOL DoW under the Ethics Requirements, SUBSOL has no intention of undertaking secondary use of personal data.

Furthermore, the interested parties will be kept informed regarding what data will be collected, stored and processed in each activity. Specifically, in those cases where personal data is involved, detailed information will be provided:

- on what personal data will be collected, stored and processed;
- on the recruitment process, inclusion/exclusion criteria for participation;
- on privacy/confidentiality and the procedures that will be implemented for data collection, storage, access, sharing policies, protection, retention and destruction during and after the project;
- on how informed consent will be pursued;
- if application/s need to be filed with a local/institutional ethics review bodies (if personal data is being collected) and if yes, which bodies/where/when.

3 Data Description

This section includes a description of the data that will be collected or generated, stored and processed in the scope of SUBSOL. The origin, nature and scale of the data will be presented as well as to whom it could be useful, and whether it underpins a scientific publication. Also information on the existence (or not) of similar data and the possibilities for integration and reuse will be included.

3.1 Collected data

First, primary and processed data from the pilot sites will be collected by the stakeholders (e.g. hydrologic, geochemical, meteorological, etc.) in order to determine the current situation and evaluate the proposed configurations.

Additional field-data sets will be collected and field visits will be performed in the sites to conduct geophysical surveys and measurements of key parameters. The nature and scale of the data collected depends on the modelling specifications as well as other possible limitations of the field work.

All these data will be properly processed to obtain the necessary information for successful SWS implementation. These data will not be kept centrally by the project (due to the large size of the respective datasets), and will remain at the custody of project partners and will be managed using data management processes of each institution. From these datasets, extracts suitably selected and processed will be included in the pilot case descriptions which will be housed in the Knowledge Base as part of the case studies descriptions. Metadata pointing to all additional data available (at partner repositories, as above) will be also included with the case studies descriptions in the Knowledge Base.

Furthermore, reports, studies and publications will be collected and processed in order to be used for the application of SWS solutions in other sites. Qualitative and other evaluation data originating from the interaction of the stakeholders will be collected and processed. All these data will be included in the project's Knowledge Base.

Possibilities for integration and reuse of similar data (e.g. from other projects and initiatives) will be explored and determined in a next project phase.

3.2 Generated data

Data generated in the scope of the project will originate from the SUBSOL monitoring stations that will be installed in the pilot sites to control important parameters. These include online real time data of water quality and water levels in the aquifers as well as the performance of treatment process in the pilot areas. Detailed water quality monitoring will allow for efficiency evaluation of geochemical reactions.

Issues addressed include selection of appropriate sensors also to monitor the performance of horizontal wells, monitoring strategies, and data-to-knowledge transfer, which should ultimately lead to integration of well field optimisation systems and SCADA/PLS in a combined control.

Additionally, data will be produced through the application of sophisticated groundwater modelling (e.g. aquifer models for groundwater management) that will be applied to enable automated control of the SWS applications.

All this data will be properly processed to secure successful SWS implementation and will be stored in the databases of the **SUBSOL Data Monitoring Systems** housed by the pilot owners.

3.3 Tools developed

All models, tools, routines developed or collected in the project and information related to them will be housed in the project's Knowledge Base, under the heading "toolbox" and will be publicly accessible if prior IP provisions don't explicitly preclude from open access (i.e. in the case of commercial software used in the project). In such a case, a suitable factsheet with inform

4 Standardisation and metadata

This section includes a reference to existing suitable standards of the discipline. The standardisation of format and metadata processes is under evaluation from the early stages of the project. Regardless, the goal is to use already well-established data format and metadata processes, make enhancements or create a basis for new ones whereas standards don't exist or don't fill SUBSOL requirements.

4.1 Metadata

Part of SUBSOL DMP is the metadata processes, the tools used to post-process the simulation results (task 3.2) as well as the archiving and sharing process through an online platform that will be linked to existing networks, including EIP on Water Market place (task 3.3).

In particular, standardisation of related to SWS experiences factsheets, focusing on hydro-technical and regulatory pre-requisites, cost-benefits, life cycle analysis and lessons learned will take place. Effort will be made to use common APIs, glossaries and semantics with the EIP Water Market Place and relevant contact between SUBSOL and the developers of the Marketplace has been initiated.

On this basis, a continuous effort is necessary in order to define standardisation requirements and harmonisation of the output data in the process of implementing SWS.

Data documentation will ensure that the data will be understood and interpreted by any user. It will explain how the data was created, what the context is for the data, structure of the data and its contents, and any manipulations that have been done to the data.

Data documentation should include information with regards to Context of data collection, Data collection methodology, Structure and organization of data files, Data validation and quality assurance, Data manipulations through data analysis from raw data, Data confidentiality, access and use conditions.

The collected should be accompanied with information like variable names and descriptions, definition of codes and classification schemes, codes of, and reasons for, missing values, definitions of specialty terminology and acronyms, algorithms used to transform data, file format and software used.

There are a variety of metadata standards, usually for a particular file format or discipline. The most prominent standards identified at this point of time include the Ecological Metadata Language[3] and the Data Documentation Initiative (DDI)[4] to document numeric data files.

[3] https://knb.ecoinformatics.org/#tools/eml

[4] http://www.ddialliance.org/

4.2 Standardisation activities

Presently there are numerous standardisation activities including risk assessment and protection of a high level of services, although sometimes optional, is considered integrated in the entire operational process.

Other related Standards include:

- ISO 9001:2015 provides an integrated approach to quality management - putting quality at the heart of business, touching on business resilience.
- ISO/IEC 27001 is an internationally recognised best practice framework for an information security management system.
- ISO/TR 37150:2014 Smart community infrastructures - Review of existing activities relevant to metrics addresses community infrastructures such as energy, water, transportation, waste and information and communications technology (ICT), focuses on the technical aspects of existing activities which have been published, implemented or discussed. Economic, political or societal aspects are not analysed.
- ISO 19115, Geographic Information – Metadata
- ISO 19110 Geographic information – Methodology for feature cataloguing
- ISO 19139, Geographic Information – Metadata -Implementation Specification
- OGC Catalog Service

Efforts will be made that all textual information collected in the SUBSOL project and especially all new information produced and stored in the SUBSOL databases will be UTF-8 encoded.

4.3 Standardisation of data sharing

Presently there is an international gap in data exchange formats for transferring SWS information between different parties. A thorough investigation has revealed a lack of common data exchange formats and this is a challenge that SUBSOL aspires to cover in a comprehensive manner.

As a first step towards this standardisation, we suggest the use of JSON objects through RESTful web services. **JSON** (JavaScript Object Notation) is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate. It is based on a subset of the JavaScript Programming Language (Standard ECMA-262 3rd Edition - December 1999). JSON is a text format that is completely language independent but uses conventions that are familiar to most programmers.

JSON is built on two structures:

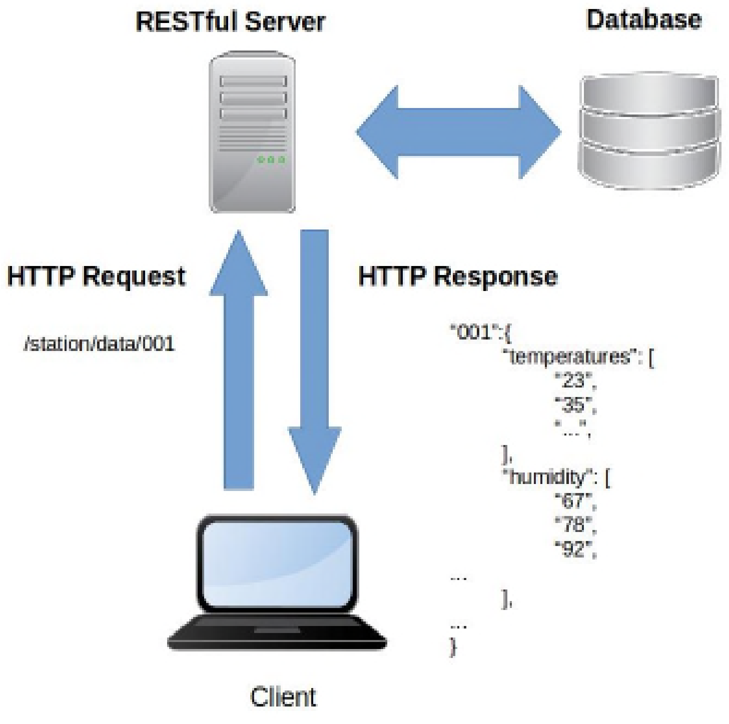
- A collection of name/value pairs. In various languages, this is realized as an *object*, record, struct, dictionary, hash table, keyed list, or associative array.
- An ordered list of values. In most languages, this is realized as an *array*, vector, list, or sequence.

These are *universal* data structures. Virtually all modern programming languages support them in one form or another. It makes sense that a data format that is interchangeable with programming languages also be based on these structures.

Additionally, a RESTful service provides a window to its stakeholders so that they can access its resources. An experienced programmer knows how to implement an easy to maintain, extend, and scale RESTful service.

The focus of a RESTful service is on resources and how to provide access to these resources. A resource can easily be thought of as a JSON object described above. A resource can consist of other resources. While designing a system, the first thing to do is identify the resources and determine how they are related to each other. This is similar to the first step of designing a database: Identify entities and relations.

Resources are available through a RESTful web service via HTTP requests. For example a request of the type "/stations/data/meteo/001" might respond with a JSON object containing all recordings of all parameters for the station provided in the request, through a standard HTTP response (see Figure 1).



```
"001":{
        "temperatures": [
                "23",
                "35",
                "...",
        ],
        "humidity": [
                "67",
                "78",
                "92",
...
        ],
...
}
```

Figure 1: An example of a REST API request to retrieve all meteorological data from a station with id 001

5 Archiving and data sharing in SUBSOL

5.1 Open data repository

It is the declared policy of the EC as far as possible to enable third parties to access, mine, exploit, reproduce and disseminate (free of charge for any user) research data that have been collected or produced as a result of research projects5. The Commission has enabled access to and reuse of research data generated by Horizon 2020 projects through the Open Research Data Pilot (ORD Pilot) in which SUBSOL participates.

Following the recommendations of the Commission the project data will be stored in a suitable repository such as **Zenodo**. Zenodo is a general-purpose open access repository recommended for research data of any kind. It assigns all publicly available uploads a Digital Object Identifier (DOI) to make the upload easily and uniquely citable. The infrastructure has been developed and is supported by CERN which guarantees data safety and availability. Zenodo is an OpenAIRE and CERN collaboration and is one of the research data repositories recommended by the EC.

The data which will primarily be uploaded to the repository are the 'underlying data', i.e. the data needed to validate the results presented in scientific publications, including the associated metadata, as well as any other data (for instance curated data not directly attributable to a publication, or raw data), including the associated metadata. The following is a list of the first data to be uploaded to the repository

- SUBSOL Knowledge Base and Marketplace data
- Dataset of the SUBSOL Data Monitoring System
- SUBSOL Taxonomy
- SWS Toolkit Knowledge Base data
- Use cases for the Online SWS Platform

5.2 Data sharing

SUBSOL will establish pilots for assessing the efficiency and impact of SWS implementation. This will serve as the primary enabler for an outreach programme intended to maximise uptake of SWS solutions in new markets in the open-source community. The pilot results, insights and methodology will be available through the online SWS Platform, whose key feature will be a virtual market place (Task 3.3) while the field data generated by SUBSOL’s monitoring program will be housed in the SUBSOL Real-time Data Monitoring System as discussed above. We will present next the project’s strategy for keeping the relevant data safe through archiving as well as the approach related to data sharing in each case.

[5] EC Directorate-General for Research & Innovation. H2020 Programme - Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020, Version 3.2, 21 March 2017

5.3 Data backup and recovery

Developing and establishing a backup and recovery plan for all SUBSOL data is a crucial part of data security. While a general purpose repository like Zenodo will store and publish the significant releases of the data produced in SUBSOL, a backup plan takes care of the daily backups of all data. A variety of backup software exist for all operating systems, enabling manual and automated backup of the data. State of the art backup systems offer efficient storage of data (full and incremental backup, differential, backwards deltas) which keep track with older versions and even allow the reconstruction of the data at any single point in time.

From the perspective of the data security, we distinguish between the following data categories:

- Data having no overall importance for the project. These data are often temporary and/or serve as an intermediate step for the production of other data. They are usually stored in various files. For this category of data it is in the responsibility of the local system administrators to establish reliable backup plans.
- Data stored in the project’s website or in the SUBSOL Knowledge Base described in sections 0 and 0 will be backed up centrally in a remote backup server. The backup server will be located physically at a different site than the other servers and will be connected with them through the Internet. Thus it will not be affected by any malfunction or damage which may occur at the site of each server. A filesystem snapshot utility based on rsync (e.g. rsnapshot) will be used to make periodic snapshots of local servers, and transfer the data securely via ssh protocol to the remote backup server.
- Data collected by the monitoring systems at the various case study sites (see section 0). These data can grow into large volumes depending on the size of the monitoring network, the number of the observed parameters and the recording time step. In case of high frequent observations, more frequent backups or establishing a replication system will be required. Depending on each case, monitoring data may be and stored in local and/or remote backup servers.
- The SUBSOL Git repository stored in a remote server will serve as a means to backup software code produced during the project (see section 0).

A backup and disaster recovery plan for all data categories will be developed at an early stage of the SUBSOL project. The plan will include the steps to be taken before, during and after a disaster as well as the backup frequency and rotation scheme. Each plan will be tested in a dry run.

6 Exceptions or additional services

At this point the project does not anticipate collecting and/or generating data that should be exempted from open access. If such a case is encountered (e.g. due to (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related or security-related concerns) the Data Management Plan will be updated accordingly.