HACID - Deliverable

Data Management Plan

This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101070588. UK Research and Innovation (UKRI) funds the Nesta and Met Office contributions to the HACID project.

Deliverable number:	D1.3
Due date:	30.11.2022
Nature[1]:	DMP
Dissemination Level[2]:	PU
Work Package:	WP1
Lead Beneficiary:	CNR
Contributing Beneficiaries:	MPG, Human Dx EU, Nesta, MetOffice

[1] The following codes are admitted:

○ R:	Document, report (excluding the periodic and final reports)
○ DEM:	Demonstrator, pilot, prototype, plan designs
○ DEC:	Websites, patents filing, press & media actions, videos, etc.
○ DATA:	Data sets, microdata, etc.
○ DMP:	Data management plan
○ ETHICS:	Deliverables related to ethics issues.
○ SECURITY:	Deliverables related to security issues
○ OTHER:	Software, technical diagram, algorithms, models, etc.

[2] The following codes are admitted:

- ○ PU – Public, fully open, e.g. web (Deliverables flagged as public will be automatically published in CORDIS project's page)
- ○ SEN – Sensitive, limited under the conditions of the Grant Agreement
- ○ Classified R-UE/EU-R – EU RESTRICTED under the Commission Decision No2015/444
- ○ Classified C-UE/EU-C – EU CONFIDENTIAL under the Commission Decision No2015/444
- ○ Classified S-UE/EU-S – EU SECRET under the Commission Decision No2015/444

Document History

Version	Date	Description	Author	Partner
0.1	16.11.2022	Creation	Vito Trianni	ISTC-CNR
0.2	09.02.2023	Initial draft	Ralf Kurvers	MPIB-MPG
0.3	15.02.2023	Contribution to Section 3	Alessandro Russo	ISTC-CNR
0.4	16.02.2023	Contribution to Sections 2 and 3	Andrea Giovanni Nuzzolese	ISTC-CNR
0.5	21.02.2023	Contribution to Sections 2, 4, 5 and 6	Vito Trianni	ISTC-CNR
0.6	22.02.2023	Contribution to Sections 1, 2 and 7	Aleks Berditchevskaia	Nesta
0.7	24.02.2023	Contribution to Section 2.3	Rita Marques	Nesta
0.8	24.02.2023	Review	All	All
0.9	28.02.2023	Final review	Vito Trianni	ISTC-CNR
1.0	08.03.2023	Final document	Ralf Kurvers	MPIB-MPG

Table of content

1. Overview

This deliverable consists of the first, preliminary data management plan for the HACID project. It outlines all the data the project already has or currently foresees it will develop, collate, or use. The plan provides an overview of the data the project partners anticipate will be collected as well as relevant data management practices.

The data management plan will be updated throughout the lifetime of the project to reflect project activities and a future iteration will be submitted at M18.

2. Data Summary

In this section, we describe the different data exploited within the project, following standard requirements. Specifically, descriptions will be provided for each activity within the project that is expected to generate relevant data to be maintained and preserved. In this respect, the identified activities corresponds to (i) the reuse/generation of knowledge bases to support research in each of the case studies, namely medical diagnostics (MD) and climate services (CS); and (ii) the user research and participatory research, which inform both case studies as well as the evaluation of collective intelligence in practice. For each activity, we will describe the planned data management practices answering to the following set of standard questions:

- *What is the purpose of the data collection/generation and its relation to the objectives of the project?*
- *What is the origin of the data?*
- *Will you re-use any existing data and how?*
- *What types and formats of data will the project generate/collect?*
- *What is the expected size of the data?*
- *To whom might it be useful ('data utility')?*

2.1. Medical Diagnostics

The medical diagnostics use case aims at testing the HACID-DSS for decision support in general medical diagnostics problems. It exploits and extends the online platform developed by the Human Diagnosis Project (Human Dx, see https://www.humandx.org/). The main purpose is to test the effectiveness and efficiency of the new HACID technology with user-provided cases. Within the medical diagnostic use case, we will compare the quality of the individual (and collective) diagnostic decisions using different implementations of the HACID technology (and different aggregation mechanisms), and compare these against baseline decisions without the HACID technology.

Data relevant for this use case are related to (i) the diagnostics cases and solutions by expert human raters on which the HACID methodologies will be tested, and (ii) the domain knowledge formalised into a knowledge graph that is used for decision support. The former are submitted via the Human Dx platform by expert users (i.e. healthcare professionals), and mainly consist of solutions to medical cases (real or fictitious), which contribute to the collective effort to provide crowd-sourced diagnoses. The latter are assembled from consolidated terminological and encyclopaedic resources from the medical domain, publicly available in online repositories.

Within the HACID project, we expect to generate new data as well as re-use existing data (i.e., already collected/generated by Human Dx) to test novel ways to aggregate decisions in medical diagnostics. Moreover, existing data may be used as a baseline to test against newly collected data (i.e. with employing HACID technology). For what concerns the re-use of terminological resources, we will rely on the SNOMED-CT knowledge base. SNOMED-CT will be re-used as an authoritative knowledge source for clinical terms, to enrich and ground the data collected and processed, and to enable the collective intelligence and data aggregation approaches that will be investigated. Finally, we also plan to reuse encyclopaedic knowledge bases such as Wikidata, as a source of structured knowledge that will be re-used for enriching the data collected and processed in project's use cases. Additional domain-specific data that will be re-used will be identified as part of project activities.

Different types of data will be exploited, often coming in heterogeneous formats. Most relevantly, data from human raters consist of diagnostic decisions, possibly associated with process-related data such as confidence and response times. Moreover, we will ask raters to evaluate the quality of diagnostic decisions in relation to the ground truth of cases for testing purposes. These data will be formatted in delimiter-separated values files (CSV and TSV data formats), with the size of datasets ranging in the order of 10MB to 100MB, thus rather small.

To formalise the domain knowledge, the project will produce and rely on semantic knowledge graphs that combine ontologies defined using the Web Ontology Language standard (OWL) and linked data defined according to the Resource Description Framework standard (RDF). Knowledge graphs will be produced from data available in different formats, from delimiter-separated values files (CSV and TSV data formats), to JSON files, to ontologies/datasets already available in OWL/RDF. The knowledge graph formalises domain knowledge on how to represent diagnostics cases and includes SNOMED-CT represented as an OWL ontology, which contains nearly 7 million triples[3]. The size of this domain knowledge graph will further increase when it will be enriched with and linked to other resources (e.g., Wikidata, as mentioned before).

The data produced within this use case may be useful to other researchers in various disciplines interested in human decision making within medical diagnostics, and in artificial and collective intelligence. Domain experts and professionals in medical diagnostics may be interested in the novel data and data infrastructures generated through the knowledge engineering part of the HACID technology.

2.2. Climate Services

This use case is tailored to develop a platform for decision support in the context of climate change adaptation management. Contrary to the medical diagnostics use case, here there is no previous experience to be exploited, and the whole technology needs to be developed from scratch. Hence, the definition of the data that will be generated and reused is a matter

[3] the number of triples (where a triple is a statement in the form of subject-predicate-object) is the most common measure of the size of a knowledge graph

of ongoing research activities (see also Section 2.3 below). However, it is already possible to predict that data will be structured in a similar way as for the medical diagnostics use case.

Data will originate from domain experts, who will provide (i) information of what structured knowledge must be represented to correctly grasp the climate service domain, (ii) judgement elicitation on relevant climate information for specific cases, and (iii) answers to specific case-based questions. In addition to structured data, the project will also deal with unstructured data, in particular textual data gathered from documents (mainly PDF files, Word documents, HTML pages), that will be processed to produce structured knowledge.

The project will re-use existing publicly-available data, which will be scraped from online repositories (e.g., scientific publications repositories such as Google Scholar or Scopus) and relevant climate data and information sources. Taxonomies and ontologies relevant for climate information and services will also be considered for inclusion, according to the ongoing needs for the decision support platform.

All the data sources will converge into a large base of domain knowledge formalised through semantic knowledge graphs, following the OWL/RDF standards, as discussed above. Knowledge graphs will then be used to produce structured dataset (e.g., CSV or TSV formatted datasets) to be used as input for specific data analytics tools and environments (e.g., statistical software, machine learning toolkits) and/or to be made available as self-contained datasets. The size of the case-specific data can vary from dozens of MB to a few GB, depending on the amount of information that will be scraped and retained from online sources. Also in this case, the knowledge graph is expected to contain a few millions triples.

The structured domain knowledge that will be produced in this use case will be relevant to climate scientists and user communities, who can benefit from the effort in combining knowledge derived either from human experts or from automatic knowledge extraction methods. It will be a unique dataset with intrinsic value, which will benefit future research and innovation in AI and climate services.

2.3. User Research

User research data will cover early activities that will help to define the design of the HACID prototypes for medical diagnostics and climate services, as well as participatory evaluation activities.

In WP6 and WP7, we will undertake user research and participatory research activities including interviews, workshops, focus groups, wiki-surveys, etc to gain qualitative insights into the key opportunities and barriers for designing and implementing the HACID technology in the medical and climate services fields. In WP5, we anticipate collecting further interview data and qualitative data from experts working in the field of participatory AI as well as project partners through structured action research activities, e.g. action learning sessions and diary studies.

Data resulting from user research will consist of interviews and surveys among selected stakeholder groups and domain experts. During participatory evaluation activities, we anticipate collecting data through video and audio recordings which will be transcribed into text data, as well as structured data through surveys.

In addition, existing user research data will be consulted where available and provided by the medical diagnosis and climate services teams (e.g., survey results, personas). We do not anticipate having to process or store this data.

All raw text, audio and video data will be held in .xlsx, .pdf, .docx, .mp3, .wav, .mp4 format on Nesta's local data storage systems. The size of all files combined should not be larger than a couple of GB.

Anonymised and interpreted data will be shared with partners. The data gathered through user research activities will be used for research only and to inform activities in other work packages. We do not anticipate publishing any of the raw data. In anonymised form, the data may be useful for designers and technologists working within the field of human-computer-interaction and participatory design.

3. FAIR data

In this section, we provide a description on how the data management practices address the principles for Findable, Accessible, Interoperable, and Reusable data (FAIR). Each section is accompanied by a set of standard questions that inform the planning.

3.1. Making data findable, including provisions for metadata

Will data be identified by a persistent identifier?

Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Will metadata be offered in such a way that it can be harvested and indexed?

Data (and metadata) will be identified by persistent identifiers. To this end, datasets will be deposited at open digital repositories (e.g., Zenodo or OSF) and we will rely on the globally unique and persistent identifiers that the repository assigns to data and related metadata, in particular Digital Object Identifiers (DOIs). Semantic assets and knowledge graphs that will be produced (including ontologies and linked data) will be entirely based on persistent, dereferencable and globally unique Uniform Resource Identifiers (URIs). To ensure uniqueness and persistence of URIs for semantic resources, we will rely on the well-established w3id.org service provided by the W3C Permanent Identifier Community Group. The service is listed in the FAIRSharing registry service[4] endorsed by the Research Data Alliance (RDA).

Upon dissemination of the results (i.e., submission or publication of a manuscript), the data which are allowed to be shared (see below) will be made available (using a persistent identifier) as an additional resource to supplement the submitted/published work.

To the extent possible (see more details below), we will make the data sets available using sufficiently rich metadata to allow accessing and understanding the data. Specifically, resources, datasets and other semantic assets produced in the project will come with rich metadata defined according to well-established metadata standards.

Resources deposited in open digital repositories will be documented using the metadata definition scheme adopted by the repository (e.g., in the case of Zenodo, metadata are

[4] https://fairsharing.org/FAIRsharing.S6BoUk

defined according to the DataCite Metadata Schema). Datasets will be described and accompanied by a set of rich metadata that document their availability and principal characteristics, thus augmenting their findability and discoverability. Specifically, the general metadata will be defined in compliance with the DCAT Application Profile[5] (DCAT-AP), based on W3C's Data Catalogue vocabulary (DCAT); where applicable, GeoDCAT-AP[6] (the geospatial extension of DCAT-AP) will be considered for geospatial datasets. Also the ontologies that will be defined in the project, as semantic assets, will be described by rich metadata defined in compliance with the Asset Description Metadata Schema Application Profile[7] (ADMS-AP).

In the repositories where the data are stored, we will use appropriate keywords relying on the metadata policy of the repository where the data is published, to allow discovery of data. Keywords/tags are, for example, part of the metadata defined in the DCAT vocabulary (dcat:keyword property) and we will provide keywords for the datasets documented using this vocabulary. To increase discoverability and potential re-use, we will also define metadata to associate our datasets with appropriate themes (exploiting the dcat:theme property) using the reference “Data theme” controlled vocabulary[8] maintained by the EU Publications Office.

For data deposited in open digital repositories we will rely on the indexing and harvesting capabilities used and offered by the repository (e.g., in the case of Zenodo, metadata are indexed by the repository itself and then searchable though its search engine; metadata are also harvestable using the OAI-PMH protocol and retrievable through a REST API).

For data allowed to be shared, the use of DCAT-AP as metadata schema for describing datasets and semantic assets will make the metadata harvestable and indexable by data portals that adopt this metadata vocabulary, including the European Data Portal.[9] This implies that, where applicable, metadata will be made available for inclusion and publication in national /international data portals and catalogues.

In the case of domain-specific ontologies that could emerge from project activities, we will also consider the possibility of having the ontologies and related metadata harvested and indexed by disciplinary thematic portals and registers (e.g., BioPortal).

3.2. Making data accessible

Repository:

Will the data be deposited in a trusted repository?

Have you explored appropriate arrangements with the identified repository where your data will be deposited?

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Anonymized data will be made accessible via trusted repositories, such as Zenodo or OSF, upon dissemination of the results, to create unique persistent URLs and DOIs, making data

[5] https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/release/211

[6] https://semiceu.github.io/GeoDCAT-AP/releases/

[7] https://joinup.ec.europa.eu/collection/support-centre/solution/asset-description-metadata-schema-adms

[8] https://op.europa.eu/s/xOGE

[9] https://data.europa.eu/en

fully accessible. GitHub will be used as repository for semantic assets (ontologies, controlled vocabularies, datasets) produced within the project, as well as for other digital artefacts that are relevant for data management (e.g., documentation and diagrams, data transformation and processing scripts, data analysis scripts, any other source code).

Data:
Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.
If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.
Will the data be accessible through a free and standardized access protocol?
If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?
How will the identity of the person accessing the data be ascertained?
Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

For the medical diagnostic use case, the proprietary data that has already been gathered or compiled by Human Dx, and that will be shared with the HACID partners to carry out the activities of the project, cannot be made openly available as is. Selected subsets of the Human Dx data will be made available as attachments to the scientific papers published by the HACID consortium. Data collected and generated within the HACID project via the Human Dx platform will be made available (in anonymized and pseudonymized form).
During the project runtime (and for two years after) there are no restrictions on the use of data by the project partners. For non-project partners, the data will be freely accessible once published.
Internal project data will be shared between the project partners through secure means, for instance as password-protected files stored on secured servers. As the project consortium is rather small, there is no need for a specific data access committee.

Metadata:
Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?
How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?
Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

Metadata will be made openly available and licenced in the public domain. In open digital repositories such as Zenodo metadata are publicly accessible and licensed under public domain. Metadata will contain information to enable the user to get access to the data, i.e., the persistent identifier assigned to the dataset by the data repository and a specification of the access conditions (e.g., open, restricted or closed).

In the case of semantic assets whose metadata will be defined using the DCAT-AP vocabulary, metadata will contain information to access the data that is both human- and machine-readable, including the landing page of a dataset distribution (`dcat:accessURL`), a direct link to a downloadable file that constitutes the dataset (`dcat:downloadURL`), and information on access rights (`dct:accessRights`) and licence (`dct:license`).

When sharing the data online, we will also include the metadata. Both will remain available and findable via platforms such as OSF and Zenodo.

We generally aim at producing data in open formats that do not require specific (proprietary) software to be accessed and read. Where relevant, information about the software needed to access or read the data (such as name of the software, download link, source code repository, etc.) will be provided as part of the documentation about the data reachable through reference(s) defined as part of the metadata.

3.3. Making data interoperable

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices?

Which ones?

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

Will your data include qualified references[10] to other data (e.g. other data from your project, or datasets from previous research)?

We will focus on data and metadata interoperability considering both the technical and the semantic perspective.

From a technical perspective, we aim at producing data and metadata using open, machine-readable, standardised formats. For structured (meta)data, we will rely on open formats that include delimiter-separated values files (CSV/TSV data formats), JSON and JSON Schema, XML and XML Schema. When using open digital repositories for our data, we will rely on the metadata formats that the repositories use for metadata representation. For instance, in the case of Zenodo, metadata is internally represented in JSON format according to a defined JSON schema; metadata can be exported in different formats following standard metadata schemes (e.g., Dublin Core, MARCXML, DataCite Metadata Schema). Knowledge graphs and semantic assets (ontologies and linked data) will be defined and produced using standard, formal, open and machine-readable knowledge representation languages from the Semantic Web stack, in particular the W3C Web Ontology Language (OWL) and Resource Description Framework (RDF). Semantic assets will be serialised and made available in different formats (considering RDF common serialisation

[10] A qualified reference is a cross-reference that explains its intent. For example, X is regulator of Y is a much more qualified reference than X is associated with Y, or X see also Y. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/)

formats). The possibility of querying and accessing knowledge graphs through SPARQL endpoints will allow retrieving (meta)data in multiple open formats (i.e., as SPARQL query results serialised in JSON, CSV/TSV, XML ecc.).

To maximise semantic interoperability, we will reuse as much as possible well-established standard data and metadata vocabularies and schemes. As already mentioned, for data deposited in digital repositories we will rely on the metadata standards and schemes used by the repositories, which in turn may rely on external vocabularies (e.g., Zenodo reuses Open Definition's concepts for licenses, Funder Registry and its metadata for representing funders, OpenAIRE's Research Graph for grants, etc.). When producing semantic metadata, we will use the DCAT Application Profile (DCAT-AP) vocabulary, based on W3C's Data Catalogue vocabulary (DCAT) which is grounded in the Dublin Core metadata standard. Metadata about the ontologies that will be defined in the project will be produced in compliance with the Asset Description Metadata Schema Application Profile (based on W3C ADMS vocabulary, which in turn builds on DCAT). When defining project ontologies, we will consider the reuse of and alignment to existing domain ontologies and other well-established vocabularies. Where applicable, our ontologies will thus be aligned with and reuse well-established vocabularies, in particular W3C's ontologies (e.g., SSN/SOSA, FOAF, ORG, SKOS, PROV-O), ISA2 SEMIC Core Vocabularies, and EU Vocabularies defined by the EU Publications Office. The possibility to reuse domain-specific ontologies and controlled vocabularies will emerge as part of project activities. At this stage, for the standardisation and re-usability of medical terminology models, we will rely on SNOMED Clinical Terms (SNOMED CT). As mentioned earlier, we also plan to re-use encyclopaedic knowledge bases such as Wikidata to increase interoperability of our data and knowledge graphs. For all other data generated we will publish relevant metadata, and consider existing schemas for any research data we publish in the future.

From a methodological perspective, we will consider the best practices defined in well-established authoritative FAIR guidelines, including the FAIR Data Maturity Model Specification and Guidelines[11] and Data.europa.eu data quality guidelines[12]. In addition, the knowledge engineering processes that we will follow in the project for ontology design and knowledge graphs production will be based on the eXtreme Design (XD) methodology, which takes interoperability into account by promoting the reuse of existing ontologies and ontology design patterns (ODPs) [1].

Where possible, ontologies or vocabularies that will be defined in the project will extend/reuse and will be aligned to existing commonly used ontologies, as outlined before (e.g., W3C's ontologies, EU's core and controlled vocabularies, etc). Alignments and mappings will be defined using the standard semantic properties defined for this purpose as part of the OWL/RDF(S) specifications (`owl:equivalentClass`, `rdfs:subClassOf`, `rdfs:subPropertyOf`, etc.), as well as the SKOS mapping properties (`skos:exactMatch`, `skos:closeMatch`, etc.). Additionally, we will design ontologies by re-using ontology design patterns [1] (ODPs) which have been demonstrated in literature to be effective solutions for fostering ontology interoperability and re-use [2] and building complex ontology network associated with large knowledge graphs [3], [4].

Differently from medical diagnostics, climate services is a relatively new discipline and thus still developing standard ontologies. Nevertheless, in this case we will design ontologies by re-using solutions such as the Semantic Sensor Network (SSN) and the Sensor,

[11] https://doi.org/10.15497/rda00050

[12] https://op.europa.eu/s/xSCX

Observation, Sample, and Actuator (SOSA) ontologies[13], which are compliant with the INSPIRE data model for enabling the interoperability of geospatial infrastructures among EU member countries. There is increasing evidence showing the effectiveness and appropriateness of using SSN/SOSA for modelling knowledge graphs in the domain of climate services [5], [6]. Hence, we will endeavour to build on existing vocabularies and ontologies, such as SSN/SOSA and will openly publish our own ontology to allow others to continue to re-use, refine and extend it. Additionally, we will provide alignments with ontologies, such as DOLCE UltraLight+DnS [7], to foster semantic interoperability from a foundational perspective. All these actions are compliant with the FAIR principles and focus on re-use and interoperability.

Semantic knowledge graphs produced in the project will build on the principles of linked data and will thus include qualified references to other data in the form of semantic links between entities/concepts. Where applicable, entities in our knowledge graphs will be linked to existing linked open datasets through standard semantic properties for entity linking (`owl:sameAs`, SKOS mapping properties such as `skos:exactMatch`, `skos:closeMatch`, etc.). Additional qualified references will be used and defined according to the ontologies that will be reused and/or defined in the project, in particular when entities from existing datasets (e.g., SNOMED-CT, Wikidata, GeoNames, etc.) will be (directly) used in the knowledge graph (e.g., to create a link between a diagnosis and a disease as defined in the SNOMED-CT vocabulary).

3.4. Increase data re-use

How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licences, in line with the obligations set out in the Grant Agreement?

Will the data produced in the project be usable by third parties, in particular after the end of the project?

Will the provenance of the data be thoroughly documented using the appropriate standards?

All documentation and artefacts needed to support data interpretation and re-use will be made available through a GitHub repository. This will include digital artefacts such as readme files with rationale and instructions, data generation/transformation/mapping scripts, data processing and analysis scripts, configurations for entity linking tools, graphical diagrams for ontologies, example queries for data exploration, ecc.

Data will be released with a clear and accessible standard data usage licence, as per Grant Agreement. The project partners aim as much as possible at making data available with an open standard licence to maximise re-use, for example by adopting the Creative Commons Attribution 4.0 International Public License (CC BY 4.0).

During and after the end of the project, data produced in the project and published will be usable by third parties in accordance with the terms of the licence under which they are released.

[13] https://www.w3.org/TR/vocab-ssn/

Quality assurance practices will follow standard approaches as from the Data.europa.eu data quality guidelines [8].

For data that feeds the semantic knowledge graphs produced in the project, we will rely on the W3C PROV Data Model and the corresponding PROV-O ontology[14] as a consolidated data model and semantic standard for representing provenance, from the simple definition of data sources to the specification of complex processes and activities that originate the data. When needed, data provenance information will thus be part of the knowledge graphs and will be expressed in a formal language and machine-readable format.

4. Other research outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.

At this point in time, there is no research output that requires management according to FAIR principles. Nevertheless, the project will produce software that will be evaluated in the context of the exploitation strategy, and for which corresponding management practices will be defined, based on established good practices such as the REUSE guidelines[15].

5. Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.) ?

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

Who will be responsible for data management in your project?

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

The responsibility of the data management resides with different project partners as various partners will work on different data. Each partner will be responsible for the data they are working on, and the responsibility is different for the different partners. There will be little additional costs incurred to make the data and other research outputs FAIR, and no additional budget is needed for this.

[14] https://www.w3.org/TR/prov-o/

[15] https://reuse.software/spec/

6. Data security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?
Will the data be safely stored in trusted repositories for long term preservation and curation?

Generally speaking, the data owners are expected to have their back-up and storage policies in place and handled by their IT departments. During the project all partners producing relevant data follow their own protocols for data storage (e.g., on institutional servers), back-up procedures, etc. By Milestone 2 (Month 9), we will investigate whether all partners have their security policies sufficiently in place or whether additional arrangements need to be made. This may especially apply to those partners working in one way or another with personal data (see also Section 7 below).

7. Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).
Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

The Ethical aspects of the HACID project—also related to data management—are described in detail in the DoA. Additionally, deliverable D1.9 Ethics Check Report has been produced substantially confirming the approach described in the DoA. Since, at this point in time, there are no deviations from the described policy envisioned, these documents remain valid. Here, we provide a brief description of the general principles that underpin ethicality of data management practices within HACID.

All questionnaires, surveys, interviews and other research activities that involve participants/stakeholders beyond project partners will include information about data management practices. Following these activities, a small amount of demographic data could be stored and processed (e.g., age and gender, profession). A privacy statement together with an explicit consent to the processing of the personal data will be distributed among data subjects, together with all the details regarding the participant rights as regulated by GDPR.

The collected data are minimised to serve the sole goals of the project. We will proceed with pseudonymisation of all the personal data, keeping separate electronic records of the data and the direct identifiers on different secure servers.

Considering that two partners are UK based, a data import/export issue could be raised. Given the planned experimentation, no data will be exported from the EU to the UK, but the collected data could be imported as some participants are recruited in the UK. However, only pseudonymised data will be imported. Participants will be informed about that, and a corresponding consent will be collected. Moreover, the import of data will be done in the frame of the recently published adequacy decision for the UK.

The procedures to import data from the UK will be conducted following standard data security practices, exploiting secure servers and password-protected access to the files. The expected amount of data to be imported is likely small (in the order of dozens of MBs). At this point in time, there is no evidence of the need for higher security approaches.

8. References

[1] A. Gangemi, ‘Ontology Design Patterns for Semantic Web Content’, in *The Semantic Web – ISWC 2005*, vol. 3729, Y. Gil, E. Motta, V. R. Benjamins, and M. A. Musen, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 262–276. doi: 10.1007/11574620_21.

[2] E. Blomqvist, V. Presutti, E. Daga, and A. Gangemi, ‘Experimenting with eXtreme Design’, in *Knowledge Engineering and Management by the Masses*, vol. 6317, P. Cimiano and H. S. Pinto, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 120–134. doi: 10.1007/978-3-642-16438-5_9.

[3] V. A. Carriero, A. Gangemi, M. L. Mancinelli, A. G. Nuzzolese, V. Presutti, and C. Veninata, ‘Pattern-based design applied to cultural heritage knowledge graphs: ArCo: The knowledge graph of Italian Cultural Heritage’, *Semantic Web*, vol. 12, no. 2, pp. 313–357, Jan. 2021, doi: 10.3233/SW-200422.

[4] A. G. Nuzzolese, A. L. Gentile, V. Presutti, and A. Gangemi, ‘Conference Linked Data: The ScholarlyData Project’, in *The Semantic Web – ISWC 2016*, vol. 9982, P. Groth, E. Simperl, A. Gray, M. Sabou, M. Krötzsch, F. Lecue, F. Flöck, and Y. Gil, Eds. Cham: Springer International Publishing, 2016, pp. 150–158. doi: 10.1007/978-3-319-46547-0_16.

[5] C. Roussey, S. Bernard, G. André, and D. Boffety, ‘Weather data publication on the LOD using SOSA/SSN ontology’, *Semantic Web*, vol. 11, no. 4, pp. 581–591, Aug. 2020, doi: 10.3233/SW-200375.

[6] J. Wu, F. Orlandi, D. O’Sullivan, and S. Dev, ‘An Ontology Model for Climatic Data Analysis’, in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, Brussels, Belgium, Jul. 2021, pp. 5739–5742. doi: 10.1109/IGARSS47720.2021.9553547.

[7] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider, ‘Sweetening Ontologies with DOLCE’, in *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, vol. 2473, A. Gómez-Pérez and V. R. Benjamins, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 166–181. doi: 10.1007/3-540-45810-7_18.

[8] Publications Office of the European Union, Data.europa.eu data quality guidelines, Publications Office of the European Union, 2022.
https://data.europa.eu/doi/10.2830/333095