D5.1

Data management plan

i~HD

Funded by the European Union

Revision v1.0

Work package	WP5
Task	5.1
Due date	31-05-2023
Submission date	31-05-2023
Deliverable lead	i~HD
Version	V1.0
Authors	Nathan Lea, Dipak Kalra (i~HD)
Reviewers	Stephen Phillips, Michael Boniface (UoS); Anna Rizzo (DW); Andoni Beristain (VICOM); Sigrid Van der Veen (UMCU); Antonio Almeida, Catalina Barrientos Gomez, José Carda, José Maria Moreira (GLSMED LH); Gustavo Hernandez (UPM)

Abstract

This deliverable provides the *Data management plan* (DMP) initial draft for SYNTHEMA. The deliverable is based on the European Commission template for Horizon 2020 projects[1]. SYNTHEMA has populated this DMP in line with recommended EC guidelines. It will be updated as the project proceeds.

SYNTHEMA is novel in that the data assets to be produced will be synthetic in nature. There may yet be a risk of re-identification given that the synthetic data will be generated based on inference from real data. The plan therefore considers this paradigm carefully in its plans.

The deliverable describes the motivations for conducting a DMP, the approach taken and the findings thus far. These preliminary findings will be updated by M24 of the project.

Keywords

Data management plan, FAIR, open data, open science, security, confidentiality, governance, compliance

[1] https://ec.europa.eu/research/participants/data/ref/h2020/gm/reporting/h2020-tpl-oa-data-mgt-plan_en.docx

Document revision history

Version	Date	Description of change	Contributor(s)
v0.1	25-04-2023	1st version of the deliverable submitted to co-authors for comment and update.	Nathan Lea, Dipak Kalra
v0.2	26-04-2023	Updates from co-author and WP5 leader	Dipak Kalra, Stephen Phillips
v0.3	24-05-2023	Series of updates and reviewer comments	Anna Rizzo, Nathan Lea, Andoni Beristain, Sigrid Van der Veen, Antonio Almeida, Catalina Barrientos Gomez, José Carda, José Maria Moreira
v1.0	31-05-2023	Final review and submission by Project Coordinator	Gustavo Hernandez

Disclaimer

Copyright notice

Document information

	Nature of the deliverable	R
	Dissemination level	
PU	Public, fully open. e.g., website	✓
CL	Classified information as referred to in Commission Decision 2001/844/EC	
SEN	Confidential to SYNTHEMA consortium and Commission Services	

*** Deliverable types:**
R: document, report (excluding periodic and final reports).
DEM: demonstrator, pilot, prototype, plan designs.
DEC: websites, patent filings, press and media actions, videos, etc.
OTHER: software, technical diagrams, etc.

Table of contents

Acronyms and definitions

AI	Artificial intelligence
AML	Acute myeloid leukaemia
DMP	Data management plan
DoA	Description of action
DOI	Digital Object Identifier
DPIA	Data protection impact assessment
EAB	Ethical advisory board
FHIR	Fast Health Interoperability Resources
GDPR	General data protection regulation
ML	Machine learning
MRI	Magnetic Resonance Images
OMOP	Observational Medical Outcomes Partnership
SCD	Sickle cell disease
WP	Work package

1 Executive summary

This deliverable provides the *Data management plan* (DMP) initial draft for SYNTHEMA. The deliverable is based upon the European Commission template for Horizon 2020 projects[2]. SYNTHEMA has populated this DMP in line with recommended EC guidelines. It will be updated as the project proceeds.

A DMP is an important component of any data intensive programme because it imposes a need for balance between protection of data, success of the programme and the potential for reuse of data.

SYNTHEMA is novel in that the data assets to be produced will be synthetic in nature. There may yet be a risk of re-identification given that the synthetic data will be generated based on inference from real data. That data may be anonymous or pseudonymised at the point of processing, but it will be detailed. The plan therefore considers this paradigm carefully in its plans.

The approach to developing the data management plan has included workshop discussions with partners at the December 2022 Kick-off meeting in Madrid and subsequent online meetings and workshops since. The details gathered were compared with the proposal and obligations of the partners as described in the Consortium Agreement.

The results of the detail gathering are presented as the DMP in Section 3 of this deliverable. It concludes with the next steps and specification of updates in time for the M24 updated version of the DMP.

[2] https://ec.europa.eu/research/participants/data/ref/h2020/gm/reporting/h2020-tpl-oa-data-mgt-plan_en.docx.

2 Introduction

2.1 Reasons for a DMP

A DMP is an important component of any data intensive programme because it imposes a need for balance between protection of data, success of the programme and the potential for reuse of data to support Open Science in adherence to the *Findable, Accessible, Interoperable and Reusable* (FAIR) principles. SYNTHEMA is a novel project because the primary data handling is focused on generating realistic and reusable synthetic data to train *Artificial Intelligence* (AI) tools that can be trained to assist in the disease management of rare diseases. Within SYNTHEMA, real health multimodal data are processed of two rare diseases, including *sickle cell disease* (SCD) and *acute myeloid leukaemia* (AML), to help develop AI models which in turn will generate the synthetic data. The goal of SYNTHEMA is to preserve the privacy of data subjects and generate realistic synthetic data. A key challenge for SYTHEMA is to balance the realistic nature of the synthetic data and reduce the likelihood of reidentifying individuals from it.

This places the development of the DMP in an unusual light because it is not only addressing **synthetic data**, but also considers the **original source data captured from patients and registry participants**. The source data is a special category of personal data, but the generated synthetic data may not immediately be classified as a special category of sensitive data and therefore should have no sharing and reuse restrictions. This, however, cannot be a foregone conclusion as there is little or no evidentiary basis for the impact on risks to the privacy and well-being of data subjects who have contributed data into the synthetic generation matrices. One goal of SYNTHEMA is to assess these risks and gather evidence as to the effectiveness of the approach in terms of privacy protection. This forms an integral part of the DMP so that these risks can be assessed and the reusability of both the original and synthetic data can be approached. In any event, the extent to which regulations such as the EU 2016/679 - GDPR[3] and the recently passed AI Act[4] will impact SYNTHEMA remains unclear so a cautious approach will be maintained.

This DMP will therefore consider both **the source data and the synthetic data**. It will further address adherence to data protection and security regulations as well as concerns around security. Data reuse must respect individual autonomy and privacy of participants. The source data that will be used to train the algorithms, will not be widely shared outside of SYNTHEMA or used for any other purpose. The synthetic data will likely be made more widely available. Whilst

[3] EUR-Lex - 32016R0679 - EN - EUR-Lex (europa.eu)
[4] EUR-Lex - 52021PC0206 - EN - EUR-Lex (europa.eu)

the DMP alone cannot provide assurance that these are addressed, it does provide a basis to inform the activities that do, including the Data Protection by Design and Default and DMP approaches. These will be documented in other deliverables.

A unique feature of SYNTHEMA is that its data processing outputs will be synthetic in nature. Whilst generated from real data relating to research participants and patients, a key goal of SYNTHEMA is to assess the faithfulness, data protection and privacy implications of the synthetic data that has been produced. To that end, the adherence to Open Science and the FAIR principles must yet be determined based on the results of the privacy assessments and wider regulatory compliance work of WP5. This in turn will rely on the analyses conducted under WP7 for ethical requirements and the appointment of an independent Ethical Advisory Board of experts.

The final decisions on particulars around sharing and onward sharing of synthetic data will be provided in M24 of the project when this DMP will be updated. This will be informed by the development of the regulatory compliance and risk assessment approaches, further understanding of the ethical challenges and development of the data pipelines and updated materials. Where items are marked as to be determined, the intention is to update these when the details have become clearer.

2.2 Approach

The DMP has been developed using the Data Protection by Design and Default principles, notably the commencement of a *Data Protection Impact Assessment* (DPIA).

A first step in the DPIA process is to establish a detailed **data flow diagram** that shows the intended data processing requirements. From this, SYNTHEMA can identify where data is originating, where it is being processed, and how and where it is being generated. Since part of the generation includes advanced processes for synthetic generation, these form part of the coverage of the DMP and feed into its development.

The information to inform the DMP has included **clarification on proposed data handling** at the kick off meeting in December 2022 held in Madrid. Following on from that, the WP5 Team have engaged with WP1 and WP2 in terms of the **clinical input and platform design** to better understand the data flows. WP5 has engaged with the clinical sites providing data to understand their **existing ethical and regulatory permissions**, and the extent to which they can share real data for further reuse if at all, and to what extent they would be prepared to share that data.

i~HD, UoS and SBA have worked very closely to study and assess the data flows for the project. This has been critical to understanding the precise details of data handling at each step of the process. The flows and a description of the governance processes will be available under D5.2. This has helped to inform the production of the DMP specified under Section 3.

3 Results – Data management plan

3.1 Data summary

Aspect	Response/explanation
Purpose of the data collection/generation and its relation to the objectives of the project	Data from existing sources is to be gathered within the data provider centres and used to **generate synthetic data that is representative and realistic across rare diseases, including SCD and AML**. The purpose of this is to assess the privacy enhancing features of synthetic data, help provide further data sources for machine learning where the population cohort is relatively small and to develop enhanced synthetic data generation.
Types and formats of data generated or collected by the project	Source data will include **clinical, imaging and omics data**. Clinical data will include laboratory results, cytology, histopathology, treatment data, outcomes and survival and limited demographics including age and gender. Imaging data will also include MRI scans, *whole slide images* (WSI) from bone marrow biopsies, 3DTI, FLAIR and DWI sequences. Omics data will include genomic data, *single nucleotide panels* (SNP) for *genome-wide association studies* (GWAS) and next-generation studies sequences (HBB, HBA1 and HBA 2 genes). Omics data will furthermore include metabolomics data. The data format will be the **OMOP *common data model* (CDM)** and will be mapped according to a harmonisation approach agreed across the work packages[5]. For the **SCD use case**, clinical data includes structured clinical data, laboratory, disease complications, time to event, response to therapy, omics data including GWAS, metabolomics, and imaging data including radiomics (*magnetic resonance imaging,* MRIs), 3D-T1, FLAIR and DWI sequences. For the **AML use case**, clinical data includes structured clinical data, laboratory, treatment information, overall survival, disease-free survival, omics data including DNA target sequencing data, cytogenetics, RNA sequencing, and imaging data including WSI histopathological images from bone marrow biopsies.

[5] Data Standardization – OHDSI

	These data items will be used to generate the synthetic data sets that will thereafter be validated for accuracy, privacy enhancement and use for expanding the data holdings for machine learning activities in the disease areas.
Any re-use existing data and how this will be done	**Existing registry and biobank data** will be used for the project and **further data collection from the data provider partners** as listed below. Reuse will be achieved within the bounds of existing consents and ethical approvals. Where necessary, amendments to existing approvals will be made if the SYNTHEMA reuse requirements are not covered under existing approvals. Prospective data collection will also occur under fully informed consent and where needed new ethics committee approvals from collection sites will be obtained.
The origin of such data	Existing data will be provided from **local data repositories at participating clinical and academic centres**, i.e., *University of Bologna* (UNIBO), *Humanitas Research Hospital* (ICH), *Vall d'Hebron Research Institute* (VHIR), *University Medical Centre Utrecht* (UMCU), *Assistance Publique Hopitaux de Paris* (APHP Henri-Mondor), and *Charité - Universitatsmedizin Berlin* (CHA). Prospective data collection will occur at the *University of Padova* (UNIPD) and *Glsmed Learning Health SA* (GLSMED LH).
Expected size of the data	In terms of population numbers across the two use cases and all sites, SYNTHEMA will process the data of approximately **3,000 patient participants**. The storage requirements are not yet clear pending the agreement on harmonisation of the data sets.
Likely users of the data	**AI and ML developers and researchers** are likely to process the data to generate and train the algorithms for synthetic data generation and validate the synthetic data on a statistical utility perspective. **Clinical researchers** will review the synthetic datasets for quality and faithfulness purposes.

Table 1. Data summary.

3.2 FAIRification of data

3.2.1 Making data findable, including provisions for metadata

Aspect	Response/explanation
Are the data produced and/or used in the project discoverable, identifiable and locatable by means of a standard identification mechanism?	The proposal will be to generate data into an agreed format, likely **OMOP CDM**. This will be labelled in accordance with the harmonisation processes for the existing data and the target models that must be agreed upon by WP1-4.
What standard identification mechanism are used (e.g., persistent and unique identifiers such as Digital Object Identifiers)?	A **project identifier** will be provided for each of the original source data items and generated data sets. These will be linkable back to the original data items that were used to generate the data. The identifiers will follow an agreed process. The format of the identifiers has not yet been decided upon but will also work alongside DOIs for the generated data.
Is meta-data available through catalogue?	A **catalogue of metadata** will be published for the synthetic data items. In the catalogue, a reference to the source data will be added. For the source data itself, the Data Controllers remain responsible for deciding whether they wish to create their own catalogue if they have not already done so.
Can meta-data be used for search?	The privacy and security risk assessments for WP5 will assess the risks of wider searching and sharing as part of the assessment scenarios. The **metadata will nevertheless be searchable and published** assuming risks to the original participants are minimal.
Naming conventions used	This has not yet been agreed upon by M6 as the common model is still under development.
Clear versioning supported?	Yes – this is a core requirement for synthetic data generation.
Additional keyword search supported?	Yes
What metadata will be created using which standards?	A complete catalogue will be developed in line with the FAIR principles.

Table 2. Making data findable.

3.2.2 Making data openly accessible

Aspect	Response/explanation
Will data be made openly available as the default?	No. A main goal of SYNTHEMA will be to assess the risks of the open availability of synthetic data. This will be achieved by conducting a DPIA that has been commenced, and using a privacy and risk assessment methodology that applies metrics to identified risks and their mitigation, all under WP5. The risks are initially focussed on maintaining the anonymity of the participants in addition to the faithfulness and the utility of the generated data. Assuming a favourable assessment, SYNTHEMA will endeavour to make available the datasets that have been generated or offer an explanation as to why this is not the case. This matter will also be addressed by the Ethical Advisory Board for the project. Source data sets will not be made available outside of any existing provisions that each provider may already have in place outside of SYNTHEMA.
Which datasets will NOT be openly available and why?	This will be determined after the risk assessments have been completed.
How will the data & meta-data be made accessible (e.g. by deposition in stated repository)?	If deemed within the bounds of manageable risks, host institutions (currently UNIBO and UPM) will provide the metadata catalogue and will seek to publish it on the project website as well as existing research repositories. These have yet to be identified.
If known repository, what arrangements explored?	None yet though the particulars will be determined in the first 18 months of SYNTHEMA.
If project-specific access, then:	
Data Access Committee	The sustainability activities of SYNTHEMA are starting to explore how to reuse offerings of synthetic data that can be shared. Assuming a favourable risk review, an independent Data Access Committee would be established under the commission of either a spinout legal entity or an existing member of the consortium. Further specifics will be considered by the risk assessment process and the results of WP5.
Any conditions for access (i.e., a machine-readable license)	To be determined within the first 24 months of SYNTHEMA in accordance with the risk assessments from WP5 and EAB reviews.
What methods or software tools will be needed to access the data?	To be finalised by Month 36 of SYNTHEMA in line with WP5 and EAB assessments.
Documentation for software	All software development will be fully documented consistent with the development of a Medical Device under the Medical Device Regulations assuming Certification may be required.

Availability of software	This will likely be published Open Source pending discussions with the development teams and sustainability work packages.
Institution and researcher vetting process/procedures - describe	All institutions participating in SYNTHEMA must have the appropriate insurance and procedures in place to conduct ethically approved research in health and care. All researchers and staff must attend routine data protection training and be familiar with Good Clinical Practice guidelines, with appropriate accreditations if they are clinical and participant facing during the research. Each institution is responsible for ensuring these aspects are in place under the terms of the Consortium Agreement and approvals to participate in the action.

Table 3. Making data openly accessible.

3.2.3 Making data interoperable

Aspect	Response/explanation
Are the data produced in the project interoperable?	They will be in line with the OMOP standards.
If not, explain why not	N/A
Data and metadata vocabularies, standards or methodologies used	OMOP CDM, possibly also the Fast Health Interoperability Resources FHIR[6] (to be defined).
Standard vocabularies used	ICD 9, 10 and SNOMED CT
Mappings from uncommon or project-specific ontologies or vocabularies to more commonly used ontologies	N/A

Table 4. Making data interoperable.

3.2.4 Increase data re-use (through clarifying licences)

Aspect	Response/explanation
Will data be available for onward data-sharing/re-use?	The risk assessments and input from the EAB will determine if this is possible and appropriate. This assessment is a core part of SYNTHEMA goals.
Approach to data licensing for onward use	Likely Open Science / Creative Commons.
Likely date for data availability for onward use	By M48 of the project.

[6] Welcome to the HL7 FHIR Foundation

Explain any restriction on date of availability	Restrictions on access to source data for SYNTHEMA researchers will be defined by each of the source data Controllers but will not be made more widely available to others outside the consortium by SYNTHEMA. Access to the synthetic data generated and any restrictions applied will be determined by the risk assessments.
Possible restrictions on onward data-sharing	Where there is a substantial risk to the wellbeing of the data subjects onward sharing of synthetic data would need to be restricted. The intention is that data will be shareable outside the scope of SYNTHEMA. No source data will be shared by SYNTHEMA, except for independent assessment and review under regulatory processes.
Data retention policy (including availability for data-sharing)	This will be in line with standard medical research retention policies, approximately 20 years across participating countries.
Description of data quality assurance processes	

Table 5. Increase data re-use (through clarifying licences).

3.3 Allocation of resources

Aspect	Response/explanation
Estimated project costs for making data FAIR	To be determined by M24 and published in the updated DMP.
Data management responsibility across the project	This will defer to WP5 for regulatory compliance and the recipient of the generated synthetic data (currently UPM and UNIBO).
Resources required for long term preservation (costs and potential value, who decides and how what data will be kept and for how long)	To be determined by M24 of the project and published in the updated DMP.

Table 6. Resource allocation.

3.4 Data security

Aspect	Response/explanation
Data security measures used (including data recovery as well as secure storage and transfer of sensitive data)	The platforms used for analysing source data will be installed at the source sites and run in line with their security procedures. The data collection platform for the synthetic data will be certified to ISO 27001 and operate under security provisions to be determined by a Joint Controller Agreement that will be prepared for month 18 of SYNTHEMA.
Where data will safely be stored (in certified repositories for long-term preservation and curation). Provide detail	This has yet to be finalised but please see response above. The particulars will be determined by the risk assessment in WP5.

Table 7. Data security.

3.5 Ethical aspects

Aspect	Response/explanation
Any ethical or legal issues that can have an impact on data sharing	Data sharing will be in line with any existing approvals that have been provided to the data sources as well as any considerations for the DPIA and risk assessments in WP5. Any direction from the EAB will be adhered to as well. WP7 has also commenced an analysis of ethical considerations, and these will be specified as design and guideline requirements in deliverables as specified below.
References to ethics deliverables and ethics chapter in the Description of the Action (DoA) – if relevant	Data protection compliance – an interim report under D5.2 will be published in M12. The privacy assessment will be published under D5.3 under D5.3 in M48. An updated DMP will be published as D5.4 in M24. Quality assurance guidelines have been published as D7.1. The Ethical design requirements will be published as D7.2 in M12 and an Ethics handbook as D7.3 in M48.
Questionnaires dealing with personal data	N/A
How is informed consent for data sharing and long term preservation sought in such questionnaires?	N/A

Table 8. Ethical aspects.

4 Conclusions

This deliverable describes initial work that has been completed to understand data handling and requirements for SYNTHEMA. It has considered the data sources, recipients and required processing, and emphasised the need to conduct the appropriate assessments for understanding privacy and data protection risks for sharing and distribution of synthetic data. The DMP will inform the Data Protection Impact Assessment which is underway.

This DMP has focused on source data that is held by data providers that relates to individuals and the synthetic data that will be generated by AI and algorithms that have been trained on the source data.

The DMP has clarified that the source data will only be used for training the synthetic data generation algorithms and will not be reused or shared more widely outside of SYNTHEMA. The synthetic data generated will be handled and shared within the bounds of the FAIR principles once risks to privacy and confidentiality are assessed to address any risk of reidentifying individuals.

Some of the responses to the DMP are to be determined. There are several activities that will help confirm the responses to these. The activities include data protection compliance and privacy risk assessments, a review of the ethical implications of the project and the appointment of an independent Ethical Advisory Board.

This DMP is therefore a preliminary plan that has outlined current plans and the basis upon which they can be developed by M24 for an updated DMP. In the meantime, the assessments will continue as planned and further updates will be provided via the deliverable structure.