WILLOW

D6.4

WILLOW

Project: Wholistic And Integrated Digital Tools for Extended Lifetime and Profitability of Offshore Wind Farms	
Project number:	1011122184
Project acronym:	WILLOW
Call:	HORIZON-CL5-2022-D3-03
Topic:	HORIZON-CL5-2022-D3-03-04
Type of action:	HORIZON - RIA
Project staring date:	01/10/2023
Project duration:	36 months

Deliverable D6.4

Data Management Plan

Due date of deliverable	Month 6
Actual submission date	2024-03-31
Organization name of lead contractor for this deliverable	CEIT
Dissemination level	PU
Revision	3.1

Authors	

Author(s)	Ainhoa Cortés (CEIT)
Contributor(s)	Christof Devrient (VUB)
	Santiago Molins (TSI)
	Jeroen Tacq (Sirris)
	Koen Robbelein (VUB)
	Valentin Chabaud (SINTEF)

DISCLAIMER

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them.

DRAFT PREPARATION			
Version	Publication date	Content	Author
0.0	13/02/2024	First Draft	Ainhoa Cortés (CEIT)

HISTORY OF CHANGES		
Version	Publication date	Change
1.0	13/02/2024	First draft of the document
2.0	26/03/2024	Second draft of the document
3.0	11/12/2024	Third draft of the document

EXECUTIVE SUMMARY

The present document constitutes Deliverable D6.4 "Data Management Plan" in the framework of the Project titled "Wholistic And Integrated Digital Tools for Extended Lifetime and Profitability of Offshore Wind Farms" (Project Acronym: WILLOW; Grant Agreement No 1011122184.).

This document has been prepared to provide the data management life cycle for the data to be collected, processed and/or generated[1] by the WILLOW project corresponding to the Open Research Data Pilot (ORDP). Hence, it will report how the management project will manage the data generated as well as how in the project it will be ensured the open access to scientific publications. The presented data management plan (DMP) explains how data will be exploited or made accessible for verification and re-use, and how data will be curated and preserved.

The management project will follow, wherever possible, an open access-open source policy. This DMP describes how research data, results, and publications will be handled during the project execution and about access policies foreseen after the project closure. Furthermore, this DMP explains the life cycle of data that will be collected, processed and disseminated during the project.

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS AND ACRONYMS

Abbreviations / Acronyms	Description
CBM	Condition-Based Maintenance
CM	Condition Monitoring
DMP	Data Management Plan
DSS	Decision Support Systems
EOC	Environmental and Operational Conditions
EPC	Engineering, Procurement and Construction
FEM	Finite Element Method
IPS	Indoor Positioning System
LCoE	Levelized Cost of Energy
OPEX	Operational Expenditures
O&M	Operation and Maintenance
PdM	Predictive Maintenance
RUL	Remaining Useful Lifetime
SHM	Structural Health Monitoring
UAV	Unmanned Aerial Vehicle
US	Ultrasound
UWB	Ultra-Wide Band
WT	Wind Turbine
WF	Wind Farm
WFO	Wind Farm Operator

1   INTRODUCTION

The WILLOW project is part of the Open Access to Scientific Publications and Research Data Program in Horizon Europe. The goal of the program is to foster access to data generated in Horizon Europe projects.

Open Access refers to a practice of giving online access to all scholarly disciplines information that is free of charge to the end-user. This way, data becomes re-usable and the benefit of public investment in the research will be improved.

The purpose of the DMP is to provide an overview of the main elements of the data management policy that will be used by the Consortium regarding to the project research data. The DMP is not a fixed document but will evolve during the lifespan of the project.

The DMP covers the complete research data life cycle of the WILLOW project. It describes the types of research data that will be generated during the project, the strategies on research data preservation and the provision on access rights. The research data should be "FAIR", that is findable, accessible, interoperable and re-usable. These principles precede implementation choices and do not necessarily suggest any specific technology, standard or implementation solution.

In the context of research funding, open access requirements do not imply an obligation to publish all the results. The decision to publish is entirely up to the grant beneficiaries. Open access becomes an issue only if publication is chosen as a means of dissemination. Moreover, open access does not affect the decision to exploit research results commercially, e.g. through patenting. The decision on whether to publish through open access must come after the more general decision on whether to publish directly or to first seek protection.

The policy for Data Sharing in the WILLOW project includes different access levels for consortium members and for external users:

- Data Sharing and Access for Project Partners: all the information and research data related to the WILLOW project will be shared through a **SharePoint** Private Site to which only project partners can access. The Trello tool will be also used as a complement to manage the activities of the project.

- Data Sharing and Access for External users (Open Access):

  - **Project Website** (https://willow-project.eu/): will be a key in supporting the project communication to the general public and the project stakeholders during the project lifetime. Furthermore, the website has a section also for uploading public data and documents.

  - **Zenodo Repository**: this repository will host the Open Access Research Data during the project lifetime.

The repository Zenodo has been chosen as the main repository to store, classify and provide Open Access to the stored data objects originated within the WILLOW project frame.

Zenodo (zenodo.org) is an open, dependable repository for all scholars, enabling researchers from all disciplines to share and preserve their research outputs, regardless of size or format. The main features of Zenodo that makes it a suitable tool for data sharing and preserving are:

- Zenodo is linked to Horizon Europe projects and all results are immediately linked to OpenAIRE and the EC portal.

- Share and link research: Zenodo provides a rich interface which enables linking research outputs to datasets and funding information. All open content is harvestable via OAI-PMH by third parties.

- Supports versioning: Via a top-level DOI you can support all the different versions of a file.

- Trusted, reliable, safe: Data is stored at CERN, which has considerable knowledge and experience operating large scale digital repositories. Data files and metadata are kept in multiple online and offline copies.

- Reviewing: Research materials can set to share with reviewers only, if needed.

2   DATA SUMMARY

The system to be created in WILLOW aims to collect different kind of data to analyze the structural health of the offshore wind turbines (critical structures). The collection of data is intended, at last, to estimate and predict the lifetime assessment of those structures according to the use cases and scenarios defined in Deliverable D1.1, and the load cases defined in Deliverable D1.2 to comply with WILLOW objectives:

- Obj1: Development of a global structural health monitoring system for the tower/transition piece, and foundations based on loads, accelerations, images, thickness losses considering fatigue, pitting corrosion and coating degradation by using physical and virtual sensors combined with machine learning techniques. This solution will permit to detect and evaluate critical failures related to cracks due to fatigue or pitting corrosion or external sources and to evaluate more accurately the loads on steel tower/transition piece and on foundations by developing AI/ML approach for automatic classification of damages.

- Obj2: Development of prognosis tools to predict the consumed lifetime (CL) and the remaining useful lifetime (RUL) of all the sub-structures within the offshore windfarm by combining SCADA and SHM data using physical models and machine learning methods. This will permit to make decision towards lifetime extension and lifetime optimisation. In parallel, population-based novelty/anomaly detection methods at wind-farm level will be developed by combining historical data considering damages of different nature (corrosion, fatigue, human intervention) and the prognosis data coming from the developed tools to identify and mitigate faulty behaviour.

- Obj3: Development of decision-making support tools for wind farm operators for smart power dispatch in curtailed conditions and O&M scheduling valid for up to 20 MW turbines. This will help to optimise the wind farm operation, deciding how much power should each turbine extract from the wind in present, near and far future in order to satisfy grid, market and lifetime constraints, and hence maximise profitability.

- Obj4: Development of a set of ambitious outreach activities to facilitate the sector uptake of the tools: Dissemination, Communication, Exploitation, Standardisation and Training activities.

The preliminary and simplified scheme of the WILLOW system is shown in Figure 1.
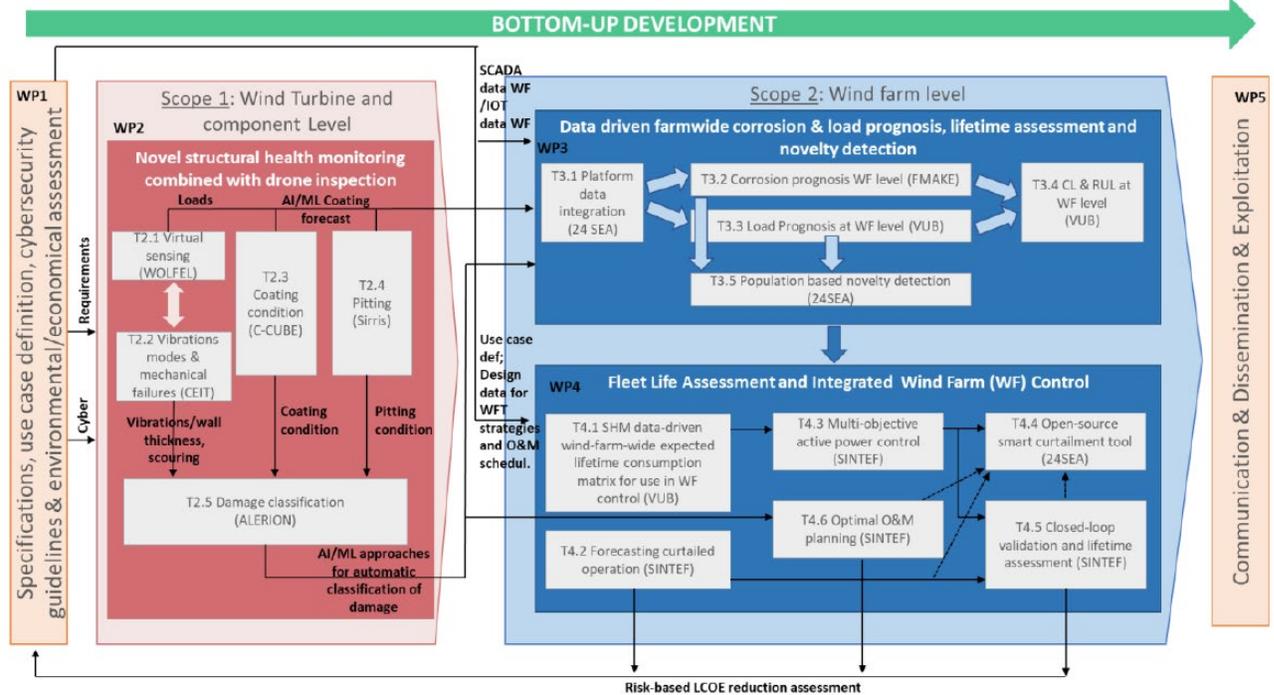
Figure 1. WILLOW preliminary system scheme

Figure 1 presents the different types of data that will be generated in the scope of the WILLOW project aligned with the presented objectives:

- Environmental data

- Coating condition data

- Corrosion testing data

- Corrosion degradation rate data

- Pitting risk data

- Damages classification data

- Corrosion prognosis data

- Load prognosis data

- Load versus corrosion data

- Maintenance scheduling data

- Wind farm fatigue data

- Wind Farm lifetime consumption data

- Wind Farm control data

In every case, the template of datasets from partners acquiring and sharing these data will cover "Origin of the data", "expected size", "format", "reusing of data".

Every dataset and working document will be properly referenced to a version control number, where major versions are attained only when sending them for a milestone of the task.

The **types and formats of data** within the project frame include the following:

- Laboratory or field data: datasets (*.txt, *.doc, *.docx, *.xls, *.xlsx, *.mat, *.json), numerical data (*.XX), qualitative data (*.txt, *.doc, *.docx), data statistics (*.xls, *.xlsx), images (*.jpg, *.png, *.jpeg, *.tiff), videos (*.avi, *.mov, animated GIF), geographical information (*.kml, *.gpx).

- Models data: statistics (*.xls, *.xlsx), graphs (*.ogg, *xls, *.xlsx), bibliography (*.enl), code and executables (*.rpm, *.exe, *.c, *.cpp, *.py, *.java, *.m)

- Scientific texts: manuscripts and reports (*.doc, *.docx, *.pdf), publications (*.doc, *.docx, *.pdf), conference proceedings (*.doc, *.docx, *.pdf), conference presentations and posters (*.ppt, *.pptx, *.pdf), books and theses (*.doc, *.docx, *.pdf).

- Dissemination material: leaflets and fact-sheets (*.pdf), images (*.jpg, *.png, *.jpeg, *.tiff), animated images (*.gif), videos (*mp4), social network publications and website (*.html),

  presentations and templates (*.ppt, *.pptx, *.pdf).

- Management documents: deliverables (*.doc, *.docx, *.pdf), patents (*.doc, *.docx, *.pdf).

In the following table, we provide more details about the data protection, data dissemination and data storage of the different type of data generated in the scope of the WILLOW project. This table will be updated in D6.5 2nd version of Data Management Plan providing for example the specific size of the data, the confidentiality and the availability after the end of the project of some sets of data.

Purpose of Data	Standards	Open Access	Confidential and why	Size	How will data be disseminated during project	How data is available after project (re-use)	Data utility	Origin/provenance of the data (Lead Partner)	Origin/provenance of the data (WP)
Cybersecurity guidelines	Models data Management data	Yes	No	To be defined	Zenodo	It will be kept in Zenodo Repository, as well as WILLOW repository	Developers, manufacturers, research community, general society	CEIT	WP2
Underwater sound at wind farm	Field data	TBD, initially the answer is no	TBD, initially the answer is yes, as data are refered to private assets belonging to a partner. Pending partner's allowance.	To be defined	To be defined	To be defined	Research community	TSI	WP1
Ultrasound sensors data (thickness loss)	Laboratory data	Yes	No	To be defined	Zenodo	It will be kept in Zenodo Repository, as well as WILLOW repository	Developers, manufacturers, research community, general society	CEIT	WP2
Ultrasound sensors data at HarshLab (thickness loss, corrosion rate, pitting quantification)	Laboratory data Models data	Yes	No	To be defined	Zenodo	It will be kept in Zenodo Repository, as well as WILLOW repository	Developers, manufacturers, research community, general society	CEIT	WP2

Purpose of Data	Standards	Open Access	Confidential and why	Size	How will data be disseminated during project	How data is available after project (re-use)	Data utility	Origin/provenance of the data (Lead Partner)	Origin/provenance of the data (WP)
Ultrasound sensors plus accelerometers data	Laboratory data Models data	Yes	No	To be defined	Zenodo	It will be kept in Zenodo Repository, as well as WILLOW repository	Developers, manufacturers, research community, general society	CEIT/TSI	WP2
Ultrasound sensors plus thermography data (damage categorisation)	Laboratory data Models data	Yes	No	To be defined	Zenodo	It will be kept in Zenodo Repository, as well as WILLOW repository	Developers, manufacturers, research community, general society	CEIT/ALERION	WP2
Ultrasound sensors plus electrochemical sensors data (corrosion, corrosion rate, pitting risk)	Laboratory data	Yes	No	To be defined	Zenodo	It will be kept in Zenodo Repository, as well as WILLOW repository	Developers, manufacturers, research community	CEIT/SIRRIS	WP2
Electrochemical sensors data at Blue Accelerator (corrosion, corrosion rate, pitting risk)	Laboratory data Models data	Yes	No	To be defined	Zenodo	It will be kept in Zenodo Repository, as well as WILLOW repository	Developers, manufacturers, research community	SIRRIS	WP2

Purpose of Data	Standards	Open Access	Confidential and why	Size	How will data be disseminated during project	How data is available after project (re-use)	Data utility	Origin/provenance of the data (Lead Partner)	Origin/provenance of the data (WP)
Spectroscopy sensors data (coating condition)	Laboratory data Models data	Yes	No	To be defined	Zenodo	It will be kept in Zenodo Repository, as well as WILLOW repository	Developers, manufacturers, research community	CCUBE	WP2
Spectroscopy sensors data at Blue Accelerator (coating condition)	Laboratory data Models data	Yes	No	To be defined	Zenodo	It will be kept in Zenodo Repository, as well as WILLOW repository	Developers, manufacturers, research community	CCUBE	WP2
Virtual sensing (load estimations at WT level)	Laboratory data Models data	Yes	No	To be defined	Zenodo	It will be kept in Zenodo Repository, as well as WILLOW repository	Developers, manufacturers, research community	WÖLFEL	WP2
Lifetime assessment	Laboratory data Models data	Yes	No	To be defined	Zenodo	It will be kept in Zenodo Repository, as well as WILLOW repository	Developers, manufacturers, research community	VUB	WP3
Load prognosis	Laboratory data Models data	Yes	No	To be defined	Zenodo	It will be kept in Zenodo Repository, as well as WILLOW repository	Developers, manufacturers, research community	VUB	WP3

Purpose of Data	Standards	Open Access	Confidential and why	Size	How will data be disseminated during project	How data is available after project (re-use)	Data utility	Origin/provenance of the data (Lead Partner)	Origin/provenance of the data (WP)
Corrosion prognosis	Laboratory data Models data	Yes	No	To be defined	Zenodo	It will be kept in Zenodo Repository, as well as WILLOW repository	Developers, manufacturers, research community	FMAKE	WP3
Novelty detection	Laboratory data Models data	Yes	No	To be defined	Zenodo	It will be kept in Zenodo Repository, as well as WILLOW repository	Developers, manufacturers, research community	24SEA	WP3
Wind Farm Control and impact on lifetime assessment and LCOE	Models data	Yes	No	To be defined	Zenodo	It will be kept in Zenodo Repository, as well as WILLOW repository	Developers, manufacturers, research community	SINTEF	WP4
Smart Curtailment tool	Laboratory data Models data	Yes	No	To be defined	Zenodo	It will be kept in Zenodo Repository, as well as WILLOW repository	Developers, manufacturers, research community	24SEA	WP4

Purpose of Data	Standards	Open Access	Confidential and why	Size	How will data be disseminated during project	How data is available after project (re-use)	Data utility	Origin/provenance of the data (Lead Partner)	Origin/provenance of the data (WP)
Operation and Maintenance scheduling	Models data	Yes	No	To be defined	Zenodo	It will be kept in Zenodo Repository, as well as WILLOW repository	Developers, manufacturers, research community	SINTEF	WP4
Progress, interim and final reports	Management documents	No	Restricted to the project partners and the EC	-	-	Website's intranet	Research community	CEIT	WP6

Purpose of Data	Standards	Open Access	Confidential and why	Size	How will data be disseminated during project	How data is available after project (re-use)	Data utility	Origin/provenance of the data (Lead Partner)	Origin/provenance of the data (WP)
Standardisation guidelines	Management data	No	Restricted to the project partners and the EC	-	-	Website's intranet	Research community Standardisation bodies	CEIT	WP5
Business cases/market studies	Templates established by partners	No	Restricted to the project partners and the EC. All subject of publications and related to the open pilot strategy will be made available	-	-	Website's intranet	Research community	BEC	WP5
Communication, exploitation and Dissemination data	Scientific texts Dissemination material	Yes	No	-	Project website, Zenodo	It will be kept in Zenodo Repository, as well as WILLOW repository	Developers, manufacturers, research community, general society	BEC	WP5

Apart from the generated data described in section 2, other data sources will be used:

-   Ceit : Uniform corrosion data generated by the EU Project WATEREYE (json file) as an input for the corrosion prognosis models (https://zenodo.org/records/13897090). This data will be used by FMAKE.

-   Norther: dataset (including information for modelling purposes (geometries and masses) and model update and validation purposes (SCADA and SHM dataset) for the full-scale validation of model-based virtual sensing methods for an operational offshore wind turbine as an input for T2.1, T2.2 and all the tasks in WP3 (https://zenodo.org/records/11093262).

-   SCADA data from Norther will also be used to calibrate and validate the power fluctuation model for wind farm control in tasks T4.3 and T4.5, in conjunction with publicly available weather hindcast data from the ERA5 and NORA3 databases.

-   Publicly available data about the European electrical grid from ENTSO-E is used for prediction of curtailment in task T4.2.

-   Sirris shared data from SOCORRO project with FMAKE, VUB and 24SEA. This concerns data on uniform corrosion rates and water quality parameters, measured at the Blue Accelerator test site. The corrosion rate data is available over a period of approximately 1.5 years, water quality parameters in two periods of a few months. The goal is to use this data to start the development of the models as outlined in WP2 and WP3, allowing us to get started while the new data to be collected as part of WP2 is not yet available. It should be noted that the SOCORRO dataset is rather limited, and doesn't include any data on coating degradation. Therefore, the new data to be generated as part of WILLOW is certainly indispensable. However, by using the pre-existing data, a headstart can be made in developing and testing some aspects of the models.

-   Sirris using data from MAXWind project. From the MAXWind project, also a limited data set of approximately one year is available on uniform corrosion rates, water quality parameters (only intermittend data) and ICCP. While this data cannot be shared with the partners due to confidiallity, it can already be used by Sirris to investigate correlations between the measured parameters. Once more data from Norther and the test at the Blue Accelerator start to become availble, the preliminary findings from the MAXWind data can be evaluated and further developed.

-   Norther shared Corrosion & ICCP data with Sirris, FMAKE and C-Cube. The data will be used in the framework of WP3 for the development of corrosion prognosis models.

3   FAIR DATA

3.1   Making data findable and re-usable

Our WILLOW website, hosted at https://willow-project.eu/ and created with the last version of CMS WordPress, will be the main tool to provide public information about the project as well as the main communication and dissemination channels for project results. Links to open datasets as well as publications (e.g., on Zenodo) will be provided on the website.

WILLOW will use Zenodo repository uploading all the scientific publications and public datasets. Any modelling code generated in the project will be hosted on the widely used platform to share version-controlled open-source computer codes, GitHub.

3.1.1   Index conventions and versioning

Project Deliverables and Publications

WILLOW_DeliverableNumber_DeliverableTitle_VersionNumber

Model-based Simulation Data and Codes

Naming conventions used in the development of codes follow common best practice rules for OOP naming, including:

- Class names shall be nouns in UpperCamelCase
- Function names shall be verbs or verb-object groups
- Variable names shall start with a lowercase letter with the exception of some common one-letter variable names like T for temperature

WILLOW Project Data (Deliverables and Publications)

- WILLOW Data will be named:
    - Deliverables: WILLOW_[DN]_[DT].[VN]
    - Publications: WILLOW_[PN]_[PT].[VN]
    - [DN]   Deliverable Number
    - [DT]   Deliverable Title
    - [PN]   Publication Number
    - [PT]   Publication Title
    - [VN]   Version Number

3.1.2   Metadata provision

Zenodo Repository offers the possibility to assign several metadata to all uploads, in order to make the content findable and re-usable. The tags offered by Zenodo will be used in this project as **metadata**:

- Publication type (journal article, presentation, book, thesis…etc.).
- Title, authors, affiliation.
- Description of the content.
- Communities that the data belong to.
- Grants which have funded the research.
- Identifiers (DOI, ISSN, PubMed ID, URLs…etc.).
- Contributors.
- References.
- Keywords.

3.1.3   Identifiability of data

Zenodo assigns all publicly available uploads a **digital object identifier** (DOI) to make the upload easily and uniquely citeable. If the upload already has a DOI assigned, it can be detailed in the metadata provision. Each data set will be tagged by a persistent digital object identifier (DOI), supplied with all relevant metadata and the WILLOW acronym and HEU grant ID.

All data generated under the WILLOW project will acknowledge the grant in the following way "*Results incorporated in this project received funding from the European Climate, Infrastructure and Environment Executive Agency (CINEA) under grant agreement Nº 1011122184*", and will automatically be associated to the project via OpenAIRE portal.

3.1.4   Keywords

All uploads will include a group of relevant keywords in order to facilitate the identification of the results.

3.1.5   Clear versioning

Zenodo Repository provides a new feature to handle versioning: DOI versioning allows to edit/update the record files after they have been published, cite a specific version of a record and cite all versions of a record.

When an upload is published on Zenodo for the first time, Zenodo registers two DOIs:

- a DOI representing the **specific version** of the record.
- a DOI representing **all versions** of the record.

Afterwards, a DOI is registered for every new version of the same upload. Each data set will be tagged by a persistent digital object identifier (DOI), supplied with all relevant metadata and the WILLOW acronym and HEU grant ID.

Versioning of datasets will follow the principles and best practices laid out by the Research Data Alliance[2]. A summary of the main versioning principles recommended by the RDA are shown below:

Revision (version control):

- A new instance of a dataset that is produced in the course of data production or data management that is different from its precursor is called a "revision".
- Any dataset revision should be identified.

Release (data products):

- The release of a new version of a dataset shall be accompanied by a description of the nature and the significance of the change.
- The significance of this change will depend on the intended use of the data by its designated user community.
- Each new release of a data product shall have a unique identifier.

Granularity (aggregates, composites, collections, and time series):

- Data may be aggregated and combined into collections or time series.
- The collection should be identified and versioned, as should each of its constituent datasets.
- Entire-time series should be identified, as should time-stamped revisions.

Manifestation (data formats and encodings):

- The same dataset may be expressed in different file formats or character encodings without differences in content. While these datasets will have different checksums, their work does not differ; they are manifestations of the same work.
- Manifestations of the same work should be individually identified and related to their parent work.

Provenance (derived products):

- The definition of revisions and releases signifies that a dataset has been derived from a precursor and is part of the description of its lineage, or provenance.
- Provenance can be more complex than following a linear path. Information accompanying a dataset release should therefore contain information on the provenance of a dataset.

Citation:

- Include information about the Release in the citation. DataCite recommends using semantic versioning, issuing a new identifier with major releases, use the "alternate identifier" and "related identifier" elements to identify releases and how they relate to other datasets, e.g., whether it was derived from a precursor

- Updating the metadata does not create a new version, it only changes the catalogue entry.

Software versioning will be done according to the principles and best practices of Semantic Versioning 2.0.0 (https://semver.org/). The summary of semantic versioning is that, given a version number MAJOR.MINOR.PATCH, the:

- MAJOR version when you make incompatible changes,
- MINOR version when you add functionality in a backwards compatible manner, and
- PATCH version when you make backwards compatible bug fixes.

Available extensions to the MAJOR.MINOR.PATCH format include additional labels for pre-release and build metadata. A full overview and further details of the Semantic Versioning Specification can be found on www.semver.org.

3.2 Making data openly accessible

3.2.1 Which data will be made openly available

Scientific Publications

All peer-reviewed scientific publications arising from Horizon Europe funding have to be made available in open access. Therefore, each beneficiary must ensure open access to all peer-reviewed scientific publications relating to its results. All the scientific publications originated by the WILLOW project will be made openly accessible by gold or green Open Access Publishing.

Other Research Data

Additionally, any other research data or information set as publishable will also be made openly accessible. However, any dissemination data linked to exploitable results will not be put into the public domain if they compromise their commercialization or have inadequate protection.

3.2.2 How the data will be made available

The open access mandate comprises two steps: depositing publications in repositories and providing open access to them.

The WILLOW project will fulfill these two steps by uploading the data to the Zenodo Repository.

All generated datasets with the tag "public" and all underlying scientific publications will be uploaded to Zenodo and accessible open and free of charge. Scientific publications and the underlying datasets will be linked through persistent identifiers. All generated data with dissemination level "Sensitive" will not be shared for commercial exploitation reasons

Metadata, including licenses for individual data records and record collections, will be harvestable by the OAI-PHM protocol using the record identifier and the collection name.

3.2.3   Methods or software tools needed to access the data

As a general rule, the format of the data deposited in Zenodo Repository will enable to access them through standard software tools like Adobe Acrobat Reader or Microsoft Office Package.

It will be possible to retrieve metadata through the public REST API. The data will be available by any web browser on www.zenodo.org. For data formats that cannot be opened using standard software tools, reliable information on the tools required to validate the results will be provided with the data.

3.2.4   How access will be provided in case there are any restrictions

As detailed in previous sections, any dissemination data linked to exploitable results will not be put into the public domain if they compromise their commercialization or have inadequate protection. In this case, the scientific committee of WILLOW will individually analyze and decide on the particular access and time restrictions for each result.

3.3   Making data interoperable

WILLOW will encourage the use of standard vocabularies for all data types present in the data sets to allow inter-disciplinary interoperability.

It's particularly important to enable the integration of diverse datasets from various sources. All documents adhere to standard scientific measures such as the S.I. units and terminologies found in ISO testing standards. Zenodo, the chosen repository for storing shared data, employs JSON schemas for metadata representation and allows for conversion into widely used formats like Dublin Core, MARCXML, BibTeX, CSL, DataCite, and integration with Mendeley. The metadata for data records will adopt Zenodo's lexicon. For specialized terminology, it refers to publicly accessible, external vocabularies, for example, licensing terms (Open Definition), funding bodies (FundRef), and grants (OpenAIRE).

This will be updated in D6.5 2nd version of Data Management Plan.

3.4   Data re-use

3.4.1   How the data will be licensed to permit the widest re-use possible

Data re-use is subjected to the license under which it is deposited on Zenodo.

All public data sets will be under the regulation of the Creative Commons Licenses. Any Data classified as sensitive will, by default, not be reusable for reasons of commercial exploitation. As default, the CC-BYSA license will be used for public data. This license allows others to remix, tweak, and build upon your work, even for commercial purposes, as long as the original work is

credited and the new work is licensed under identical terms. This license is often compared to "copyleft" free and open-source software licenses. Hence, any derivatives of the original work will also allow commercial use.

Regarding public deliverables, the associated data will be shared once the EC has approved the deliverable. In case of scientific journal publications, underlying data will be made available at the same time the journal publishes the article. The underlying data will be linked to the publication.

The Project Management Board will decide on the specific license that applies to each data deposited, taking the exploitability of the results into account.

3.4.2   When the data will be made available for re-use.

The data will be available for re-use immediately after deposition on Zenodo Repository.

Scientific publications will be uploaded to Zenodo as soon as they are published by the editorial and, at the latest, six months after publication.

Other research data not linked to scientific publication will be uploaded to Zenodo following the instructions of the Project Management Board (PMB).

3.4.3   Third parties and re-usability

The data uploaded to Zenodo, as they are deposited on a free-access base, can be re-used by third parties.

3.4.4   Data quality assurance processes

Data quality assurance is performed by the Zenodo Repository. In particular, two independent MD5 checksums are stored for each file uploaded. One checksum is stored by Invenio and is used to detect changes to files made from outside of Invenio. The other checksum is stored by EOS and is used for automatic detection and recovery of file corruption on disks. On top of that, the quality assurance of data is achieved by the chain of approval for data release. The first QA is done by the party who generate the data. Afterwards, the data are submitted to the WP leader for approval and sent to the technical committee to determine whether IPR issues exist. Finally, it is sent to the PMB for the final approval or deny.

3.4.5   Length of time for which the data will remain re-usable

All the files uploaded to Zenodo will remain re-usable for the lifetime of the repository. In case of closure of Zenodo, the best efforts will be made in order to preserve the data in an alternative repository.

4 OTHER RESEARCH OUTPUTS

No other research outputs are foreseen.

5   ALLOCATION OF RESOURCES

5.1   Costs for making data FAIR

The costs for making data FAIR are mainly those related to the cost of Open Access to Scientific Publications, as the use of Zenodo Repository is free of charge. The costs of data management activities are limited to project management costs and will be covered by the project grants. Resources needed to support data reuse after the active project period will be solved from case to case.

5.2   Responsibilities for data management in WILLOW project

CEIT is responsible for data management within WILLOW. Any member of the Consortium can upload contents in the repository.

5.3   Costs and potential value of long-term preservation

Long term preservation of WILLOW open access research data will be based on Zenodo Repository, which is free of charge. The longevity of the data curation is only limited by the lifetime of the Zenodo depository, which currently has an experimental program defined for the next 20 years at least. Self-archiving or so-called 'green' Open Access will be also applied through the developed Zenodo repository. As also required, open access to the publication will be ensured in a maximum delay of 6 months.

The Project Management Board of WILLOW will decide on long term preservation of Research data associated to exploitable results. This will be done during the project lifetime and based on the protection strategy followed by the consortium.

6   DATA SECURITY: STORAGE AND TRANSFER

All data exchanges will be performed throughout web services based on HTTPS protocol (SharePoint, CMS WordPress, and Zenodo fulfil this condition).

Regarding data security, we must take special considerations for personal data:

- Data processing must be authorized by the subject and they will be executed fairly and lawfully. In the event of detected alterations or unauthorised disclosure, the data subject will be immediately informed.

- It is forbidden to process personal data revealing more information than needed for the specific purpose.

- The data subject will have the right to remove consent, on legitimate grounds, to the processing of data related to her/him.

6.1   SharePoint

Microsoft uses some of the strongest, most secure encryption protocols in the industry to provide a barrier against unauthorized access to data. When data is at rest two types of encryptions are used: disk encryption and file encryption. On disk encryption level, a BitLocker is used to secure data and on file encryption level every file is secured with its own key that uses Advanced Encryption Standard (AES) with 256-bit keys, which is a Federal Information Processing Standard (FIPS) 140-2 compliant. Metadata for each dataset will include:

- File name
- File type
- Description
- Date
- Version
- WP number Responsible person
- Lead

6.1.1   Data Storage

SharePoint Server provides two backup systems: farm and granular.

- The farm backup architecture in SharePoint Server starts a SQL Server backup of content and service application databases, writes configuration content to files and also backs up the Search index files and synchronizes them with the Search database backups.

- The granular backup and export architecture make use of Transact-SQL queries and export calls. Granular backup and export are a more read-intensive and processing-intensive operation than farm backup. From the granular backup system, you can back up a site collection or export a site or list.

SharePoint will be activated until the end of the WILOW project. Once the deactivation is requested, all personal data are immediately locked and stored for up to 30 calendar days (in case of accidental deletion). After that, all personal data will be permanently deleted.

6.2   CMS WordPress:

The WILLOW webpage is hosted at https://willow-project.eu/, created with the last version of CMS WordPress. We will upload to the webpage the public reports and the public meta-data from our developments. These data will be stored in the databases of ARSYS, which is a domain service provider. The storage will be totally secure, and the items will be retained for the lifetime of the domain and databases.

https://www.arsys.net/infrastructure/security

More information on how CMS WordPress store our data is available here:

https://wordpress.org/about/privacy/

- **Google Analytics**: allows you to analyze the visitors of the WILLOW website. An account of Analytics and Search Console have been created and associated to the WILLOW email (analitica@willow-project.eu). Thus, we will be capable of analyzing the number of the website visitors, their locations, how many times they view or download the public documents or the public data sets.

   More information on how Google Analytics stores our data is available here:

   https://secureprivacy.ai/how-to-be-gdpr-compliant-with-google-analytics/

   https://www.termsfeed.com/blog/privacy-policy-google-analytics/

6.3   Zenodo

6.3.1   Data Storage

All files uploaded to Zenodo are stored in CERN's EOS service in an 18 petabytes disk cluster. Each file copy has two replicas located on different disk servers.

For each file two independent MD5 checksums are stored. One checksum is stored by Invenio and used to detect changes to files made from outside of Invenio. The other checksum is stored by EOS and used for automatic detection and recovery of file corruption on disks.

Zenodo may, depending on access patterns in the future, move the archival and/or the online copy to CERN's offline long-term tape storage system CASTOR in order to minimize long-term storage costs.

EOS is the primary low latency storage infrastructure for physics data from the Large Hadron Collider (LHC) and CERN currently operates multiple instances a total of 150+ petabytes of data

with expected growth rates of 30-50 petabytes per year. CERN's CASTOR system currently manages 100+ petabytes of LHC data which are regularly checked for data corruption.

Invenio provides an object store like file management layer on top of EOS, which oversees e.g. version changes to files.

6.3.2 Metadata Storage

Metadata and persistent identifiers in Zenodo are stored in a PostgreSQL instance operated on CERN's Database on Demand infrastructure with 12-hourly backup cycle with one backup sent to tape storage once a week. Metadata is in addition indexed in an Elasticsearch cluster for fast and powerful searching. Metadata is stored in JSON format in PostgreSQL in a structure described by versioned JSONSchemas. All changes to metadata records on Zenodo are versioned and happen during database transactions.

7   ETHICS

No ethics or legal issues that can impact the data sharing are foreseen. However, ethical and legal aspects are included in the prepared questionnaires prepared for WILLOW partners to collect all the needed information about the generated research data (see Annex C). On the other hand, in questionnaires dealing with personal data, informed consent for data sharing and long-term preservation will be always included.

8   CONCLUSIONS

This deliverable presents how research data, results, and publications will be handled during the project execution and informs about access policies foreseen after the end of the project. Furthermore, this DMP explains the life cycle of data that will be collected, processed and disseminated during the project.

9   REFERENCES

[1] "Open Data, Software and Code Guidelines," [Online]. Available: https://open-research-europe.ec.europa.eu/for-authors/data-guidelines/.

[2] J. Klump, L. Wyborn, M. Wu, R. Downs, A. Asmi, and J. Martin, "Principles and best practices in data versioning for all data sets big and small," 2020.

Annex A: Principles of Zenodo

Source (Date: November 2024): http://about.zenodo.org/principles/

Best Effort Principles

Zenodo does not sign SLAs (service-level agreements). This is not a weakness, it is by design and marks a philosophy that we believe is most appropriate for Science. Instead, Zenodo is run by leading practitioners according to best practices.

What Science needs is inherent reliability, or more accurately demonstrated reliability based on open best practices. Furthermore, the users should be able to influence these best practices. In the long-term, a service which is trusted is much more valuable than one for which assurances must be bought.

Service failure can never be undone. Enforcing an SLA means being prepared to litigate against the contract, which means compensation, frequently assessed on the basis of loss of revenue… but none of these concepts have any place or relevance in the free exchange of research results!

Living by these principles, Zenodo strives to make available architecture, implementation, practices and statistics. Please see for example the infrastructure page. We are also aiming to have these certified.

FAIR Principles

FAIR Principles definition as referenced from: Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016).

To be Findable:

F1:   (meta)data are assigned a globally unique and persistent identifier
- A DOI is issued to every published record on Zenodo.

F2:   data are described with rich metadata (defined by R1 below)
   • Zenodo's metadata is compliant with DataCite's Metadata Schema minimum and recommended terms, with a few additional enrichements.

F3:   metadata clearly and explicitly include the identifier of the data it describes
   • The DOI is a top-level and a mandatory field in the metadata of each record.

F4:   (meta)data are registered or indexed in a searchable resource
   • Metadata of each record is indexed and searchable directly in Zenodo's search engine immediately after publishing.
   • Metadata of each record is sent to DataCite servers during DOI registration and indexed there.

To be Accessible:

A1:   (meta)data are retrievable by their identifier using a standardized communications protocol
   • Metadata for individual records as well as record collections are harvestable using the OAI-PMH protocol by the record identifier and the collection name.
   • Metadata is also retrievable through the public REST API.

A1.1: the protocol is open, free, and universally implementable
   • See point A1. OAI-PMH and REST are open, free and universal protocols for information retrieval on the web.

A1.2: the protocol allows for an authentication and authorization procedure, where necessary
   • Metadata are publicly accessible and licensed under public domain. No authorization is ever necessary to retrieve it.

A2:   metadata are accessible, even when the data are no longer available
   • Data and metadata will be retained for the lifetime of the repository. This is currently the lifetime of the host laboratory CERN, which currently has an experimental programme defined for the next 20 years at least.
   • Metadata are stored in high-availability database servers at CERN, which are separate to the data itself.

To be Interoperable:

I1:   (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- Zenodo uses JSON Schema as internal representation of metadata and offers export to other popular formats such as Dublin Core or MARCXML.

I2:   (meta)data use vocabularies that follow FAIR principles
- For certain terms we refer to open, external vocabularies, e.g.: license (Open Definition), funders (FundRef) and grants (OpenAIRE).

I3:   (meta)data include qualified references to other (meta)data
- Each referenced external piece of metadata is qualified by a resolvable URL.

To be Reusable:

R1:   (meta)data are richly described with a plurality of accurate and relevant attributes
- Each record contains a minimum of DataCite's mandatory terms, with optionally additional DataCite recommended terms and Zenodo's enrichments.

R1.1: (meta)data are released with a clear and accessible data usage license
- License is one of the mandatory terms in Zenodo's metadata, and is referring to an Open Definition license.
- Data downloaded by the users is subject to the license specified in the metadata by the uploader.

R1.2: (meta)data are associated with detailed provenance
- All data and metadata uploaded is traceable to a registered Zenodo user.
- Metadata can optionally describe the original authors of the published work.

R1.3: (meta)data meet domain-relevant community standards
- Zenodo is not a domain-specific repository, yet through compliance with DataCite's Metadata Schema, metadata meets one of the broadest cross-domain standards available.

Annex B: Zenodo.org General Policies

Source (Date: November 2024): http://about.zenodo.org/policies/

Content

- **Scope:** All fields of research. All types of research artifacts. Content must not violate privacy or copyright, or breach confidentiality or non-disclosure agreements for data collected from human subjects.
- **Status of research data:** Any status is accepted, from any stage of the research lifecycle.
- **Eligible depositors:** Anyone may register as user of Zenodo. All users are allowed to deposit content for which they possess the appropriate rights.
- **Ownership:** By uploading content, no change of ownership is implied and no property rights are transferred to CERN. All uploaded content remains the property of the parties prior to submission.
- **Data file formats:** All formats are allowed - even preservation unfriendly. We are working on guidelines and features that will help people deposit in preservation friendly formats.
- **Volume and size limitations:** Total files size limit per record is 50GB. Higher quotas can be requested and granted on a case-by-case basis.
- **Data quality:** All information is provided "as-is", and the user shall hold Zenodo and information providers supplying data to Zenodo free and harmless in connection with the use of such information.
- **Metadata types and sources:** All metadata is stored internally in JSON-format according to a defined JSON schema. Metadata is exported in several standard formats such as MARCXML, Dublin Core, and DataCite Metadata Schema (according to the OpenAIRE Guidelines).
- **Language:** For textual items, English is preferred but all languages are accepted.
- **Licenses:** Users must specify a license for all publicly available files. Licenses for closed access files may be specified in the description field.

Access and Reuse

- **Access to data objects:** Files may be deposited under closed, open, or embargoed access. Files deposited under closed access are protected against unauthorized access at all levels. Access to metadata and data files is provided over standard protocols such as HTTP and OAI-PMH.
- **Use and re-use of data objects:** Use and re-use is subject to the license under which the data objects were deposited.
- **Embargo status:** Users may deposit content under an embargo status and provide and end date for the embargo. The repository will restrict access to the data until the end of the embargo period; at which time, the content will become publically available automatically.
- **Restricted Access:** Users may deposit restricted files with the ability to share access with others if certain requirements are met. These files will not be made publicly available and sharing will be made possible only by the approval of depositor of the original file.

- **Metadata access and reuse:** Metadata is licensed under CC0, except for email addresses. All metadata is exported via OAI-PMH and can be harvested.

Removal

- **Revocation:** Content not considered to fall under the scope of the repository will be removed and associated DOIs issued by Zenodo revoked. Please signal promptly, ideally no later than 24 hours from upload, any suspected policy violation. Alternatively, content found to already have an external DOI will have the Zenodo DOI invalidated and the record updated to indicate the original external DOI. User access may be revoked on violation of Terms of Use.
- **Withdrawal:** If the uploaded research object must later be withdrawn, the reason for the withdrawal will be indicated on a tombstone page, which will henceforth be served in its place. Withdrawal is considered an exceptional action, which normally should be requested and fully justified by the original uploader. In any other circumstance reasonable attempts will be made to contact the original uploader to obtain consent. The DOI and the URL of the original object are retained.

Longevity

- **Versions:** Data files are versioned. Records are not versioned. The uploaded data is archived as a Submission Information Package. Derivatives of data files are generated, but original content is never modified. Records can be retracted from public view; however, the data files and record are preserved.
- **Replicas:** All data files are stored in CERN Data Centres, primarily Geneva, with replicas in Budapest. Data files are kept in multiple replicas in a distributed file system, which is backed up to tape on a nightly basis.
- **Retention period:** Items will be retained for the lifetime of the repository. This is currently the lifetime of the host laboratory CERN, which currently has an experimental programme defined for the next 20 years at least.
- **Functional preservation:** Zenodo makes no promises of usability and understandability of deposited objects over time.
- **File preservation:** Data files and metadata are backed up nightly and replicated into multiple copies in the online system**.**
- **Fixity and authenticity:** All data files are stored along with a MD5 checksum of the file content. Files are regularly checked against their checksums to assure that file content remains constant.
- **Succession plans:** In case of closure of the repository, best efforts will be made to integrate all content into suitable alternative institutional and/or subject based repositories.

Annex C – Data Questionnaire

Partner

Name	Company/Institute/University Acronym	Date

Data Summary

Data Type – short description, its origin (method, software, …) and relation to the project objectives

Data Format – File type? Expected file size range? Expected number of files?

Data Utility – Will the data be re-used? How? To whom might it be useful? Partners? Externals?

FAIR Data Principles

Making Data Findable – Is metadata generated? Keywords? Naming conventions? Version numbers?

Making Data Accessible – Which methods, software or documentation are needed to access the data? Is it possible to provide the relevant method/software/documentation? Also for externals?

Access restrictions – Can the data be shared? To Partners? To Externals? Specify contractual, legal and voluntary restrictions! If data cannot be shared with externals, are there alternatives (similar data, reduced resolution, …)?

Making Data Interoperable – Is data exchange and re-use between institutions, countries, etc. easy? Access by open software tools? Common metadata vocabularies, standards? If not, do you provide mappings to common standards?

Increase Re-Use – When will the data be made available? Delay, e.g. to seek patents? If yes, why and how long

Data storage

Storage properties – Version control? Access control? Password? Data recovery? Secure transfer?

Storage lifetime – how long is it intended that the data remains available/re-usable after the project end?

Storage costs – Are there resources available for data long-term preservation? (eligible HEU costs)

Other Issues

Ethical aspects – Any ethical or legal aspects that can have an impact on data sharing

Comments – Anything, I've forgotten to ask.