VerSatilE plug-and-play platform enabling remote pREdictive mainteNAnce

Grant Agreement No	:	767561
Project Acronym	:	**SERENA**
Project Start Date	:	1st October 2017
Consortium	:	COMAU S.p.A.
		Finn-Power Oyj
		VDL Weweler BV
		WHIRLPOOL EMEA SpA
		Kone Industrial Ltd
		Engineering Ingegneria Informatica S.p.A.
		OCULAVIS GmbH
		SynArea Consultants S.r.l.
		DELL EMC
		Laboratory for Manufacturing Systems & Automation
		Fraunhofer Gesellschaft zur Förderung der angewandten Forschung
		VTT Technical Research Centre of Finland Ltd
		TRIMEK S.A.
		Politecnico Di Torino

Title	:	Data Management Plan – Initial version
Reference	:	D7.5
Dissemination Level	:	PU
Date	:	2018-03-31
Author/s	:	LMS
Circulation	:	EU/Consortium

Summary:

The SERENA project aims to develop powerful solutions to aid manufacturers in easing their maintenance burdens.

This deliverable describes the initial approach for the Data Management Plan as developed through the first months M01-M06 of the project.

Contents

List of Abbreviations

AGA	:	Annotated Model Grant Agreement
DMP	:	Data Management Plan
DoA	:	Description of Action
KPI	:	Key Performance Indicator
OA	:	Open Access
ORDP	:	Open Research Data Pilot

List of Figures

Executive Summary

The **SERENA** project participates in the ORDP. As such, the current document describes the initial version of its Data Management Plan as this was developed through the first period M01-M06 of the project.

The deliverable outlines the handling of research data that will be generated during and after the lifetime of **SERENA**. The possible ways of archiving and management of the data through available web-based platforms will be investigated. Furthermore, online databases for storing research data have been examined and the most suitable was selected to be used both by the consortium partners as well as from interested people/organizations from outside the project.

1 Introduction

Computer applications have multiple data sources defined depending on the supported functionalities and their purpose. Source data constitute a valuable source of information. Data sources can be a database, a dataset, a spreadsheet or even hardcoded data.

Although raw data, often mentioned as source data, have the potential to become information, meaning useful digital information for a specific application and purpose, it requires selective extraction, organization, analysis and formatting for presentation. Once processed data may reveal valuable information and characteristics of their origin or even enable certain predictive analytics forecasting, for example, future trends. Thus, it becomes clear that the acquisition, preservation and proper management of data may enable more efficient data-driven decision making approaches for companies, forecasting, analysis of their current practices, and identification of potential bottlenecks as well as the verification of scientific and commercial published research results.

SERENA tackles with the acquisition of raw machine/sensor data and their analysis towards enabling predictive analytics aiming towards forecasting potential failures of the equipment. Such identification may result in appropriate predictive maintenance operations to take place, while an early failure identification may additionally result in a more effective scheduling of the production operation with respect to a predictive maintenance plan, thus, reducing the overall production cost.

During the lifetime of the **SERENA** project, various types of raw data will be generated through the different pilot cases. These data will contain both machine and sensor data. In addition, datasets will be generated through the intermediate processing steps of the **SERENA** systems such as KPIs for machine's condition evaluation and/or training datasets for the machine learning algorithms.

1.1 Purpose of the DMP

A DMP typically contains information on how data are created outlining the steps for sharing and preserving them. In the context of the H2020, a DMP details what kind of data will the project generate, whether and how they will be exploited or made accessible for verification and reuse and how they will be managed and preserved [1].

This particular document has been created in order to present and analyze the first steps towards the creation of the **SERENA** project DMP. An investigation of the data needed for the various developed sub-systems within the project is ongoing and their formats and prerequisites are under examination.

In addition, the deliverable focuses on the available web-based solutions for archiving, accessing and preserving Project's data made publicly available. At this point, it should be stated that the data to be made available to the public audience will be first examined for confidentiality issues and if possible made anonymous.

1.2 Objectives and tasks of WP7

WP7 aims at the creation of impact through the dissemination of the project results as widely as possible making them known to all relevant stakeholders, maximizing at the same time the exploitation of the project's results to the benefit of the **SERENA** partners.

WP7 is appropriately structured into tasks that focus on achieving the above objectives:

- Task 7.1: is the task that focuses on the establishment of the project's web portal intended for the communication with the public, in order to effectively disseminate the project's results.
- Task 7.2: is the task that obtains the activities concerning the dissemination of projects results to scientific community and industry.
- Task 7.3: focuses on the exploitation of the project's results with respect to the background and foreground IPR policies and the respective articles of the Grant Agreement.

The consortium of the **SERENA** project acknowledges that impact may be created through knowledge circulation and innovation. Making data publicly available is recognized by the members of the consortium as well as by the European Commission as an effective approach towards innovation in the public and private sectors. As a result, an approach for the DMP of the **SERENA** system that will be introduced and developed during the project, is presented in the following sections. The confidentiality of the data will be examined too, as well as the prerequisites for archiving, making them anonymous and preserving them.

1.3 Background of the DMP

The DMP specifications are governed by the “Open access to research data” article (article 29.2) of the AGA [2]. As such, the guidelines and rules are defined on open access to scientific peer-reviewed publications and research data that all beneficiaries have to follow in projects funded or co-funded under Horizon 2020 programme.

In the context of research and innovation, OA includes providing online access to scientific information free of charge and reusable [3]. Scientific information can be:

1. Research data, meaning data used in publications, curated data and/or raw machine/sensor data.
2. Peer-reviewed scientific articles which have been published in a journal.

1.3.1 Open access to peer-reviewed journal

Open access provided by journals is called “gold” open access while open access delivered by repositories is called “green” open access. Both terms are used by the OA community focused on how OA is implemented. Gold stands for publications made available directly from the publisher while Green means that a version is available somewhere else, such as a repository. However, there are several dimensions in OA including the following:

- Rea der rights
- Reuse rights
- Copyrights
- Author posting rights
- Machine readability
- Publishing costs
- Peer review

In both cases, open access to publications and/or research data is a decision of the grant beneficiaries and not an obligation. The main points towards ensuring OA to research data and publications in the context of the **SERENA** project is illustrated in Figure 1, which has been adapted from [3].

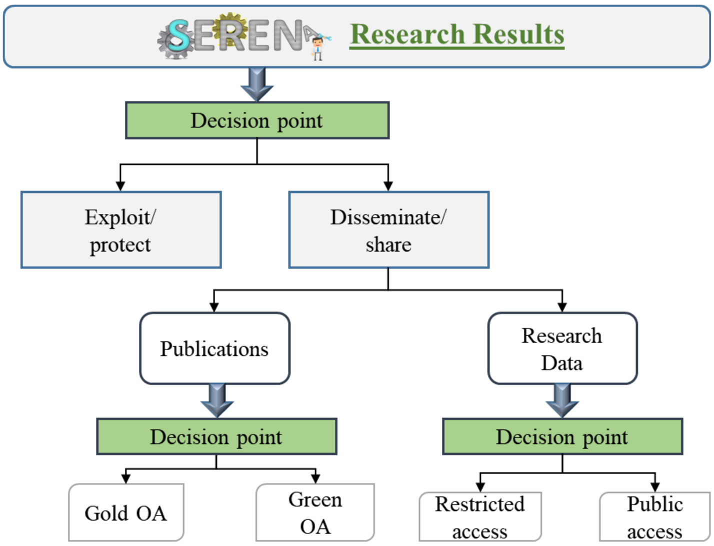


Figure 1: SERENA OA approach for data sets and publications

1.3.2 Open access to research data

Apart from publishing to an open access journal, self-archiving to an institutional repository such as INDIGO [4], or a repository supported by the EC, such as ZENODO [5], or other like the re3data repository [6], could be an option towards making something publicly available. In fact, making data publicly available is more related to making science open, which may enable the following benefits:

1. Effective scientific practices include a level communicating the evidence and validating the results.
2. Open data practices have enabled breakthroughs in certain areas of research such as crystallography, Earth observation, DNA sequencing, AI, especially when data could be reused.
3. As a result, open data may accelerate discovery through the reuse of data from the academic system and others.

2 Guiding principles

This deliverable is a living document, which will be updated regularly during the lifetime of the project. The intention of the DMP is to describe numerical models and/or datasets collected or created within **SERENA** during the runtime of the project following the guiding principles of Annex 1 as well as of the FAIR original policies [7].

Due to the fact that the project started in October 2017, there is no dataset generated or collected by the time of the compilation of this deliverable. The datasets to be made publicly available will deliver information considering the following:

- **Dataset reference and name**: Identifier for the data set to be produced. In order to be able to identify and distinguish each data set, unique object identifiers will be assigned.
- **Dataset description**: Descriptions of the data that will be generated or collected, the description element includes its types (text, spreadsheets, software, models, images, movies, audio, etc.), source (human observation, laboratory, field instruments, experiments, simulations, compilations, etc.), volume (volume of data, number of files, etc.), data and file formats (non-proprietary formats, used within community).
- **Standards and metadata**: Reference to existing suitable standards of the discipline, such as Dublin Core. If these do not exist, an outline on how and what metadata will be created. Metadata helps to categorize, understand and interpret data and may provide details about experimental setup as well as facilitate identification and discovery of new data. Metadata also tunes the data that is suggested to users.
- **Data sharing**: Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be wide open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating, in particular, the type of repository (institutional, standard repository for the discipline, etc.). In case the dataset cannot be shared, the reasons for this should be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).
- **Archiving and preservation** (including storage and backup): Description of the procedures that will be put in place for the long-term preservation of the data. An indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered.

The information listed above reflects the current concept and design of the individual work packages. The information follows a specific template and will be updated by the project-partners responsible for the different datasets to be created. With respect to the FAIR data principles [8], an initial version of a dataset template to be used for making data FAIR in an automatic approach is included in Appendix A: XML template, while a description of each field is provided in the following section.

3 Data Management related to SERENA

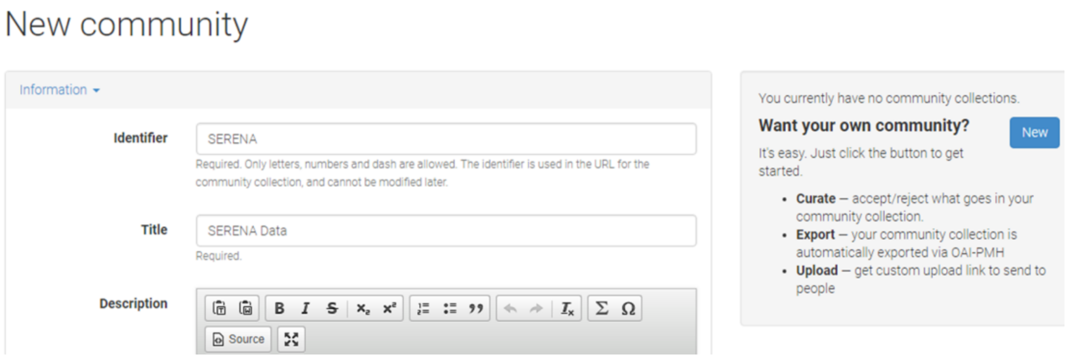
The **SERENA** consortium recognizing the importance of making research data available and easily reusable is participating in the Open Research Data Pilot in Horizon 2020. As such and with respect to section “2.2.5 Management of Data” of the DoA, observational data consisting of sensor and machine recordings have started to be collected, but are not available at the time of compiling this document. As a result, the data format is constrained to the raw data format of each source. After the collection of those datasets, from their source, the conversion to the suitable data format will be defined along with the appropriate sharing format. In order to facilitate the retrieval and reuse of the related dataset appropriate metadata values will be defined and integrated, after resolving any confidentiality issues that may be raised by the data provider. Towards this direction anonymization approaches will be considered. Datasets that will be decided to become publically available will be following the dataset format that is presented in section 3.2 of this document. Apart from sensor data, the consortium will evaluate during the development stage of the several **SERENA** components making publicly available through the channel described in the following section additional experimental data.

3.1 DMP Platforms introduction and documentation

For the **SERENA** project, the Zenodo platform has been selected for the data which will be decided by the members of the consortium to become publicly available. All research outputs from the entire scientific field can be stored in the particular platform, such as publications, posters, presentations, images and videos/audio.

A first trial account for **SERENA** project purposes was created in Zenodo. After the profile is registered and the account is activated the user can easily upload and manipulate his data files. The profile constitutes an example profile in order to serve the presentation of the platform installation to the needs of the project. A space or community for the **SERENA** project has been established, named **SERENA** Data under the following link: https://zenodo.org/communities/serena/edit/.

One of the main aspects that the platform offers is the creation of the aforementioned communities. Communities imply the dedicated storage space for a defined entity. This entity could be from research project to any other scientific procedure which demands data storage for archiving and reuse purposes Figure 2.



New community

Information ▾

Identifier: SERENA

Required. Only letters, numbers and dash are allowed. The identifier is used in the URL for the community collection, and cannot be modified later.

Title: SERENA Data

Required.

Description

You currently have no community collections.

Want your own community? New

It's easy. Just click the button to get started.

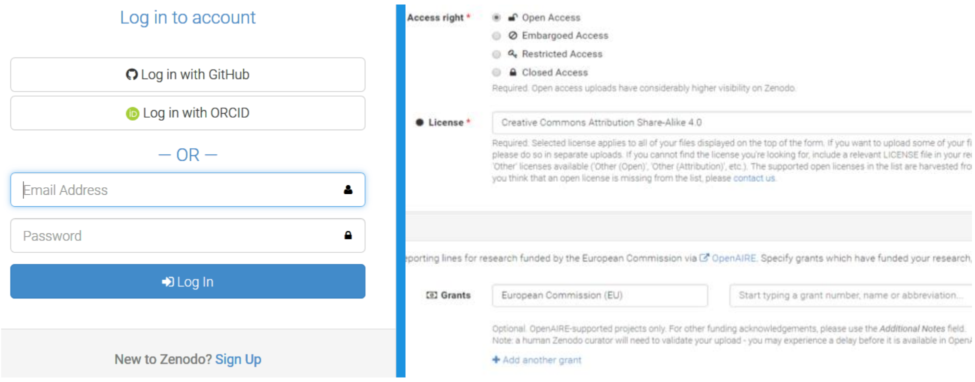
- **Curate** – accept/reject what goes in your community collection.
- **Export** – your community collection is automatically exported via OAI-PMH
- **Upload** – get custom upload link to send to people

Figure 2: SERENA project community creation in ZENODO

After the creation of the community, the creator or administrator may access it and proceed to any of the following options:

1. view the uploaded contents,
2. manage them, and
3. export the datasets

Moreover, any user with access to the community link may either search and download content or upload new datasets. In order to upload new datasets the creation of a new account is required or use of an existing one from GitHub or ORCID. In order to download pre-existing files, no registration is required. Furthermore, it provides the user with the option of uploading to establish the access rights of the files. Four types of access rights can be selected as it is depicted in, depending on the confidentiality of the data. License type can be configured in the relative tab as well as funding related information to be provided Figure 3.



Log in to account

Log in with GitHub

Log in with ORCID

— OR —

Email Address

Password

Log In

New to Zenodo? Sign Up

Access right *
- ● Open Access
- ○ Embargoed Access
- ○ Restricted Access
- ○ Closed Access

Required. Open access uploads have considerably higher visibility on Zenodo.

License *: Creative Commons Attribution Share-Alike 4.0

Required. Selected license applies to all of your files displayed on the top of the form. If you want to upload some of your fi
please do so in separate uploads. If you cannot find the license you're looking for, include a relevant LICENSE file in your re
'Other' licenses available ('Other (Open)', 'Other (Attribution)', etc.). The supported open licenses in the list are harvested fro
you think that an open license is missing from the list, please contact us.

eporting lines for research funded by the European Commission via OpenAIRE. Specify grants which have funded your research,

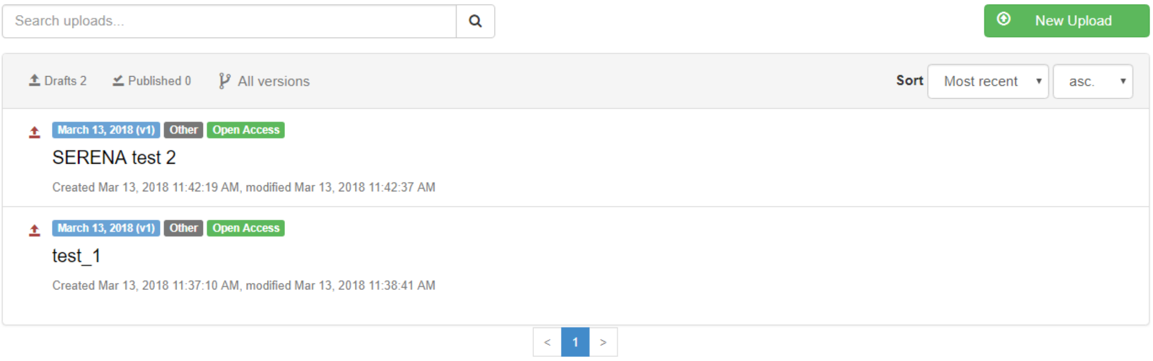
Grants: European Commission (EU) | Start typing a grant number, name or abbreviation...

Optional. OpenAIRE-supported projects only. For other funding acknowledgements, please use the *Additional Notes* field.
Note: a human Zenodo curator will need to validate your upload - you may experience a delay before it is available in Open/

+ Add another grant

Figure 3: Log in screen, access rights and license options

Two example files have been uploaded to the community of the **SERENA** Data, for which the Creative Commons Attribution-Share – Alike 4.0 has been selected. The type of the license can be reconfigured depending on the terms of each suggested license and the confidentiality level of the data Figure 4.



Search uploads...

New Upload

Drafts 2 · Published 0 · All versions

Sort: Most recent · asc.

March 13, 2018 (v1) · Other · Open Access

SERENA test 2

Created Mar 13, 2018 11:42:19 AM, modified Mar 13, 2018 11:42:37 AM

March 13, 2018 (v1) · Other · Open Access

test_1

Created Mar 13, 2018 11:37:10 AM, modified Mar 13, 2018 11:38:41 AM

< 1 >

Figure 4: SERENA data community uploaded test files

3.2 Dataset template description

As mentioned in section 2, an initial dataset template in the form of XML has been created for storing data. The XML format suggested can automate in the future the upload of data to ZENODO through a mechanism that will consume the XML and take all the required info from the XML elements. Such a mechanism can make the upload and manipulation of data a very efficient procedure and will be investigated in the future. A short description of the main data field elements included in the template is provided in the table below.

Table 1: XML elements

ELEMENT NAME	PURPOSE
SERENA_subject	The root name of each datasets referring to the **SERENA** community
datasetID	A unique identifier of the dataset
datasetDescription	A textual description of the dataset
sharingOptions	It included the sharing options of the **SERENA** subject, embargo periods, licenses, etc.
origin	It defines the main source of the dataset, such as machine name

volume	It includes the size of the dataset in MBs or GBs
Date	The date element includes the initial upload and any modification date. Furthermore, it contains a reference element which can be further linked to any other element of the XML.
contents	Under the content element, multiple elements may be included such as images, videos, documents (doc, docx, docm, pdf, ppt, etc.) as well as raw data either as plain text or in another format such as odt.
standards	This field defines any incorporated mechanism for encoding the specific dataset, along with the organization and the description of the standard.
metadata	The metadata element contains additional information over the dataset including the total number of downloads, the times that the dataset has been parsed, the ranking of the **SERENA** subject as well as the last time it was updated.

4 Conclusions

In conclusion, the requirements imposed on **SERENA** with regards to granting OA to research data have been discussed. The adopted online platform for archiving and preserving research data under the guiding principles of Annex 1 has been described. Additionally, the first steps towards populating the newly created **SERENA** data community has been made by means of two test files. The responsible partner for the archiving of the data in the proposed online platform will manage the in time update of the aforementioned tables in order for the consortium to be kept updated on the project outcomes.

This DMP includes also an XML schema to upload in a formatted and structured approach the datasets that are intended to become publicly available. Last but not least, OA regarding publications have been discussed, however for the publications, papers or deliverables, as well as for the data that are not made anonymous or are confidential the project portal will be used.

At this point the **SERENA** consortium has initiated the process of collecting data from the pilot cases. However this process also considering the confidentiality policies and data sharing restrictions of each company will require additional time. In the next stages of the project and under the decision of the responsible companies, datasets consisting primarily of sensor data may be uploaded and managed by the responsible partners according to the guidelines described in this document.

References

[1] European Commission. H2020 Programme. Guidelines on FAIR Data Management in Horizon 2020. 2016;

[2] European Commission. AGA – Annotated Model Grant Agreement. 2017;(October): 1–750.

[3] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm

[4] http://indigo.uic.edu/

[5] https://zenodo.org/

[6] https://www.re3data.org/

[7] https://www.force11.org/node/6062

[8] Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3: 160018.

Appendix A: XML template

```
<?xml version = "1.0" encoding="UTF-8" standalone="yes"?>
<SERENA_subject>
    <datasetID>
        <id>..</id>
    <datasetID/>
    <dataSetDescription>
        <description> ... </description>
    </dataSetDescription>
    <sharingOptions>
        <options>
            <option>...</option>
        </options>
    </sharingOptions>
    <origin>...</origin>
    <volume>...</volume>
    <Date>
        <upload>...</upload>
        <modified>..</modified>
        <reference>..</reference>
    </Date>
    <contents>
        <documents>
            <document>
                <filename>...</filename>
                <type>...</type>
                <size>...</size>
            </document>
        </documents>
        <images>
            <image>
                <imagename>...</imagename>
                <type>...</type>
                <size>..</size>
            </image>
        </images>
        <videos>
            <video>
                <name>...</name>
                <type>..</type>
                <size>..</size>
            </video>
        </videos>
        <rawData>
            <raw>
                <name>...</name>
                <type>...</type>
                <size>...</size>
                <source>..</source>
            </raw>
        </rawData>
    </contents>
    <standards>...
```

```
        <standard>
            <name>..</name>
            <organisation>..</organisation>
            <description>..</description>
        </standard>
    </standards>
    <metadata>
        <downloadsNo>..</downloadsNo>
        <clicks>..</clicks>
        <ranked>..</ranked>
        <lastUpdateDate>..</lastUpdateDate>
        <tags>
            <tag>..</tag>
        </tags>
    </metadata>
</SERENA_subject>
```