D1.2 DATA MANAGEMENT PLAN (LAST MODIFIED: 31-01-2020)

1. GENERAL DESCRIPTION OF DATA

1.1 What kinds of data is your research based on? What data will be collected, produced or reused? What file formats will the data be in? Also give a rough estimate of the size of the data produced or collected?

In this research project quantitative measurements data will be collected via sensor networks.
Benchmark models to generate data for specific industrial processes will be also used.
Existing and well-established datasets suitable for benchmarking the algorithms to be researched in the project, will also be used. One such example is the Tennessee Eastman process dataset.
Interviews with experts in specific test cases may be conducted.

The format of the raw datasets in total will necessarily be varied: the Tennessee datasets as well as some other benchmark data are in comma-separated-values format (CSV), which is a standard format recognizable by most software tools for data processing, reporting, visualization, and advanced analytics (Excel, Tableau, Power BI, MySQL workbench, etc.) Excel file format (xlsx) will also be supported, as will traditional SQL database dumps (SQL files) and JSON documents that are supported by a variety of technologies including NoSQL databases such as MongoDB. ML-based formats such as arff will also be supported.

1.2 How will the consistency and quality of data be controlled?

The Project and Technical Coordinators will ensure the data quality based on expert evaluation related to the specific processes to be evaluated. Experiments, source-codes and data collection will the documented in detail. Processed data files are reviewed by the coordinators and the WP leaders.

2. ETHICAL AND LEGAL COMPLIANCE

2.1 What ethical issues are related to your data management, for example, in handling sensitive data, protecting the identity of participants, or gaining consent for data sharing?

Proposed research will not involve any ethical issues.

2.2 How will data ownership, copyright and IPR issues be managed? Are there any copyrights, licences or other restrictions that prevent you from using or sharing the data?

Ownership, copyright and Intellectual Property Right (IPR) issues should be agreed according to the consortium agreement signed by the partners.

3. DOCUMENTATION AND METADATA

3.1 How will you document your data to make them findable, accessible, interoperable and reusable for you and others? What kinds of metadata standards, README files or other documentation will you use to help others understand and use your data?

Data will be recorded in text files using UTF-8 format so to guarantee that is readable in most computers. Data will be generated using computer simulation from MATLAB,python, R, and Java, while the codes will be commented and documented following conventions and programming styles. Data generated via experimental settings will use IDA for data storage, the metadata model can be seen in Etsin: "The metadata description should include at least the

permanent identifier and timestamp of the metadata, the identifier of the dataset and timestamp, dataset name, subject, language, various actors, and dataset use information". [Dublin Core](#) is a metadata model that is often used. Documents and reports will be primary generated using Latex and stored in PDF format. The Latex code also follows the TeX coding conventions and programming styles and stored in text-file format (txt or tex). The documented source-codes, when applicable, will be openly available via repositories commonly used in communications engineering research (e.g. GitHub, GitLab, ???) and will contain README files.

4. STORAGE AND BACKUP DURING THE RESEARCH PROJECT

4.1 Where will your data be stored, and how will they be backed up?

During the research project the research data will be managed and stored to Zenodo service and made openly available there. Protection and back up of the data will be managed by the CERN Data Centre who is the Zenodo service provider. For ease of sharing and work during the project, data should also be hosted on a private GitLab repository hosting also the source codes developed in the project.

4.2 Who will be responsible for controlling access to your data, and how will secured access be controlled?

Data will be available for the project's researchers and cited in publications. Other researchers may be provided access after contact the Project Coordinator (all data and source codes shall be available for reproduction), who will discuss it in the Steering Group meeting and verify if the request complies with the Consortium Agreement. Stored research data should be backed up frequently.

5. OPENING, PUBLISHING AND ARCHIVING THE DATA AFTER THE RESEARCH PROJECT

5.1 What part of the data can be made openly available or published? Where and when will the data, or their metadata, be made available?

The data will be available at the Zenodo service whenever it is possible based on the Consortium Agreement. Similarly, pre-print versions of papers will be uploaded in non-profit on-line platforms (e.g., Zenodo, ArXiv or TechRxiv) and/or the project website.

5.2 Where will data with long-term value be archived, and for how long?

The solution of the long-term preservation is still under development nationally in Finland and in Europe. The preservation solutions selected will be specified during the research project.

6. DATA MANAGEMENT RESPONSIBILITIES AND RESOURCES

6.1. Who will be responsible for specific tasks of data management during the research project life cycle? Estimate also the resources (e.g. financial, time and effort) required for data management.

Researchers' time has been allocated in the proposed budget to cover the costs of preparing data and documentation for archiving (data collection, generation of results and publications). The consortium has enough available storage for the data that will be generated and will cover the expenses for storage and maintenance of the database.