STREAMLINE

WP1
Management and dissemination

D1.3

Data management plan

Expected date: 14/05/2020

STREAMLINE has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870313

PROJECT DETAILS

PROJECT ACRONYM

STREAMLINE

GRANT AGREEMENT NO:

870313

START DATE

15/11/2019

PROJECT TITLE

Sustainable research at micro and nano X-ray beamlines

THEME

H2020-INFRADEV-2018-2020 Development and long-term sustainability of new pan-European research infrastructures

DELIVERABLE DETAILS

WORK PACKAGE ID

WP1

WORK PACKAGE TITLE

Project management, communication, dissemination and exploitation

WORK PACKAGE LEADER

Jean SUSINI

DELIVERABLE ID

D1.3

EXPECTED DATE

14/05/2020

DELIVERABLE TITLE

Data management plan

DELIVERABLE DESCRIPTION

Data management plan - Task 1.2

PERSON RESPONSIBLE FOR THE DELIVERABLE

Gary ADMANS

NATURE

☐ **R- Report** ☐ **P - Prototype** ☐ **D - Demonstrator** ☒ **O - Other**

DISSEMINATION LEVEL

☒ **P - Public**

☐ **PP- Restricted to other programme participants & EC:**

☐ **RE – Restricted to a group**

☐ **CO – Confidential, only for members of the consortium**

REPORT DETAILS

VERSION

1

DATE

14/05/2020

NUMBER OF PAGES

9

DELIVERABLE REPORT AUTHOR(S)

Gary ADMANS

FOR MORE INFO PLEASE CONTACT

Gary ADMANS

STATUS

☒ **Template** ☐ **Draft**

☐ **Final** ☐ **Released to the EC**

Contents

INTRODUCTION

This deliverable is the initial Data Management Plan (DMP) for the STREAMLINE project. It is deliverable 1.3 of work package 1.

EXECUTIVE SUMMARY

STREAMLINE is a project with a single beneficiary, the ESRF. Activities within STREAMLINE conform to activities within the usual operation of the ESRF as a user facility. STREAMLINE is not expected to produce scientific data. Nevertheless, STREAMLINE will produce scientific results related to data acquired during the testing of new services implemented by STREAMLINE. This data will be collected and processed following the ESRF Data Policy. Experimental data will be collected on inert materials, and there will not be any human data subjects, i.e. patients, or research participants. As with all experimental data collected at the ESRF, following an embargo period, it will be made available through the ESRF's data server, and identifiable by using a DOI reference. Other types of data, generated by STREAMLINE will be stored and backed up on ESRF intranet servers. Certain software projects will be made open source and distributed via external servers.

STREAMLINE

STREAMLINE is an H2020 supported project with the ESRF as the single beneficiary. STREAMLINE will complement the ESRF-EBS upgrade, which has made the ESRF the first fourth-generation high-energy storage ring synchrotron in the world. With the objective to exploit the properties of the X-ray source in an optimal way and to ensure the long-term sustainability of the facility, STREAMLINE will address the key aspects of the operation of a large-scale user facility. In this context, STREAMLINE will modernise business and user operation procedures, pilot new access modes and new services, and permit higher experimental turnover by exploiting automation at the beamlines. It will introduce new tools for the benefit of industrial and academic users alike.

1. Data Summary

STREAMLINE's purpose is to make the ESRF facility more sustainable. Its focus is on improving the facility's operational model, capitalising on the new and improved ESRF-EBS X-ray source, and making possible new access modes, including mail-in of samples, through improvements to software and systems of the user office. Concomitantly, improvement of operation of the beamlines will be carried out to permit higher throughput through improvements to the sample environments, and software for beamline control, data collection and data analysis. These projects will involve a significant programming effort, with code being stored in private ESRF repositories. Many of ESRF's existing software projects are open source, and an eventual goal will be to release mature software to the light sources community (in particular via LEAPS[1]) when this is feasible, meaning when the code developed is a stand-alone system or compatible with other open source applications. Source code developed will be in text format. Software may be developed from scratch or extended from previous projects. Data analysis software is made available by the ESRF and its collaborators through the PaNdata software catalogue[2], with open source software projects hosted on GitHub.com.

New access modes and improved facilities at the beamlines will be tested through pilot user experiments that will involve X-ray data collection on inert samples such as protein crystals or energy storage materials. Personal data on humans will not be collected or treated within STREAMLINE. The X-ray data formats used is dependent on a particular beamline or instrument. There is an ongoing action to standardise the formats at the ESRF, which will be discussed in section 2.3.

A system will be developed for reporting to stakeholders through a software application that will analyse data from ESRF user experimental reports and journal publications is also under development. The data will include user experiment reports and freely available journal publication data compiled by the ESRF User Office and the ESRF library. External sources of information may include publication data available openly on the internet or via subscription to databases held by publishers, but again only accessing publication data that is available in journals.

Table 1 summarises the data types, their volume, repositories and utility.

Data type	Volume	Repository	Utility/benefit
Experimental X-ray data (various types and formats collected by the approx. 40 different beamlines at the ESRF)	Terabytes to Petabytes	ESRF data archive	Scientific community
Source code, private	Megabytes	ESRF intranet repository	ESRF User Community
Source code, public	Megabytes	Github.com	Light source community
Microsoft Office documents, photos and videos	Gigabytes	ESRF intranet document server or repository	ESRF User Community; Management and promotion of the STREAMLINE project
PDF documents	Gigabytes	ESRF intranet document server; EU H2020 portal.	Reports on STREAMLINE project to EU.

Table 1. Summary of data produced in STREAMLINE

[1] LEAPS – the League of European Accelerator-based Photon Sources: leaps-initiative.eu

[2] PaNdata Software Catalogue: software.pan-data.eu

2. FAIR data

2.1. Making data findable, including provisions for metadata

The ESRF participates in the European Project PaNOSC (Photon and Neutron Open Science Cloud: www.panosc.eu) whose aim is to make FAIR data[3] a reality for the members of its community of large European Infrastructures. An overview of the ESRF data strategy is presented in **Figure 1**.

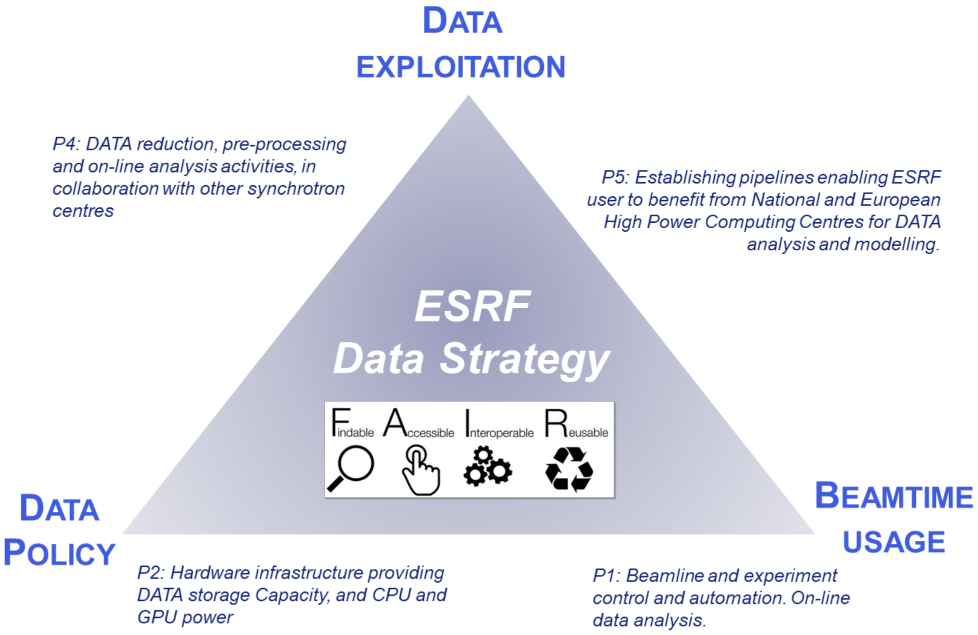
DATA EXPLOITATION
P4: DATA reduction, pre-processing and on-line analysis activities, in collaboration with other synchrotron centres
P5: Establishing pipelines enabling ESRF user to benefit from National and European High Power Computing Centres for DATA analysis and modelling.
ESRF Data Strategy
Findable Accessible Interoperable Reusable
DATA POLICY
P2: Hardware infrastructure providing DATA storage Capacity, and CPU and GPU power
BEAMTIME USAGE
P1: Beamline and experiment control and automation. On-line data analysis.

Figure 1. *The ESRF data strategy follows FAIR principles, effected through the data policy, governing data collection and exploitation.*

Data collected at ESRF beamlines are governed by the data policy[4] of the institute, which has been described in a publication[5]. The ESRF is the custodian of the raw data and its associated metadata collected at the institute's beamlines. Data is subjected to an embargo period of three years, during which time only the experimental team have access to the data. After the embargo period, which can be shortened by the owner, the data will be released under a CC-BY licence[6], with open access to anyone who has registered with the institute's data portal. Data generated on beamlines implementing the data policy (work is ongoing to render all compatible) automatically have a DOI created for them so that this identifier can be cited in publications.

Data and metadata collected at beamlines will follow the default data types and naming conventions chosen for each beamline or end station. The Nexus data format[7] has been adopted by the facility and work is ongoing to make all beamlines compatible with this standard.

[3] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.,* The FAIR Guiding Principles for scientific data management and stewardship, *Sci Data* **3,** 160018 (2016). doi.org/10.1038/sdata.2016.18.

[4] ESRF data policy: www.esrf.eu/files/live/sites/www/files/about/organisation/ESRF%20data%20policy-web.pdf

[5] R. Dimper, A. Götz, A. de Maria, V.A. Solé, M. Chaillet & B. Lebayle (2019) ESRF Data Policy, Storage, and Services, *Synchrotron Radiation News*, 32:3, 7-12, DOI: 10.1080/08940886.2019.1608119.

[6] Creative commons licence: creativecommons.org/licenses

[7] www.nexusformat.org

For the ESRF, the metadata will be stored in the ICAT metadata catalogue[8], which can be accessed online (data.esrf.fr) to browse and download metadata. An overview of data collection and storage at the ESRF is presented in **Figure 2**.

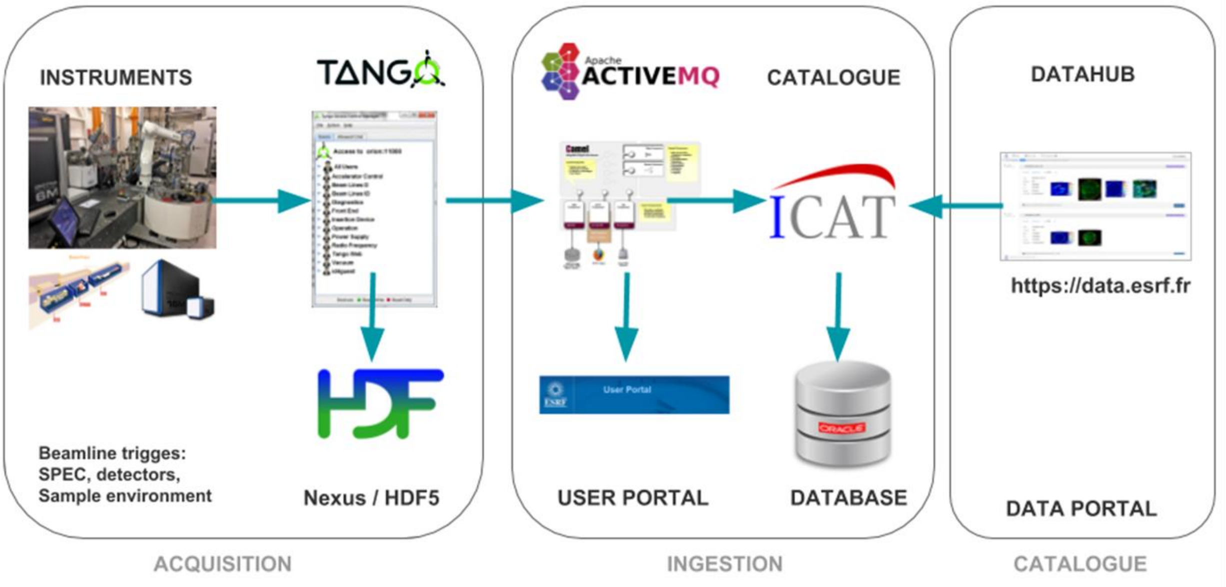

Figure 2. *Experimental data and metadata flow at the ESRF (R. Dimper et al., Synchrotron Radiation News, 32:3, 7-12, DOI: 10.1080/08940886.2019.1608119).*

For documents, the STREAMLINE programme will use an internal naming convention based on "Streamline-Dx.y_YYYYMMDD_version" e.g. "Streamline-D1.1_20200513_draft.docx".

2.2. Making data openly accessible

Data collected in the test experiments will be stored in the repositories detailed previously. The data can be made public and accessible immediately at the request of the principal investigator, but the default is a three-year embargo period, which can also be extended if needed, for example while publications are being prepared or patents being applied for.

Data in the ESRF data repository is available in formats that can be reused easily by experts in the relevant fields involving the X-ray experiments. The data can be accessed by open source software that is readily available. In the case of structural biology, new protein structures will be published in readily accessible and open databanks such as the PDB, with a structure code reference to be included in associated publications to aid other scientists to access the data.

The source code of the software developed in STREAMLINE aims to be made open source whenever feasible and published in searchable repositories such as github.com. For data analysis software, this will be publicised to the user community via the PaNdata software catalogue. Source code will remain private to the ESRF in the case where the source code extends existing closed source projects that cannot be made open source for historical reasons, such as the effort required to convert the software into generic versions that could be useful to other projects.

[8] ICAT metadata catalogue: icatproject.org

2.3. Making data interoperable

The ESRF beamlines produce a variety of X-ray data formats for historical reasons. To make data interoperable, the facility is working to make the beamlines produce raw data in the NeXus format, based on the open HDF5 format[9]. This format was designed for sharing and reuse and is vendor independent. It allows metadata to be kept with the data. From the NeXus website, "NeXus is developed as an international standard by scientists and programmers representing major scientific facilities in Europe, Asia, Australia, and North America in order to facilitate greater cooperation in the analysis and visualisation of neutron, X-ray and muon data".

Documentation generated to manage the project and its reports will be in MS Word and PDF formats which are widely readable.

2.4. Increase data re-use (through clarifying licences)

Once the embargo period has lifted, X-ray data will be published via the ESRF data repository with a Creative Commons licence. Data will be available to anyone with an account for the ESRF data repository and creation of an account is free and open to all. The default embargo period is three years. The embargo period can be shortened by the owner at any time, or extended on request when there is need to give time to publish or apply for patents. Data will be preserved for five years minimum, aiming for at least ten years.

Source code will be open source whenever feasible, and published to an accessible repository such as Github.com in plain text format.

3. Allocation of resources

Data storage is a service provided by the ESRF to its users for data collected at the beamlines. The cost of this service is included in the overall functioning costs of the institute, which are funded by the member states of the institute. The institute is curator of the data and responsible for its management and preservation.

4. Data security

Data is curated by the ESRF in order to provide data security for the researchers collecting data at the facility. The ESRF operates a very high quality of disk and tape archival with built-in redundancy using professional commercially-sourced systems. Archival systems are physically located in secure locations with access only to authorised computing staff. Data are stored on disk for mid-term storage, and backed up to long-term tape archives, which aim to preserve the data for a duration of ten years.

For source code, GitHub is considered suitable for long term storage, however, it would be prudent to store a backup copy of the source code on ESRF servers which are backed up using a long-term tape storage system maintained by the ESRF.

Word documents and pdf files generated for the administration of STREAMLINE will be backed up using the ESRF's intranet document repository.

[9] www.hdfgroup.org/solutions/hdf5

5. Ethical aspects

The data collected within STREAMLINE only concerns materials such as proteins and energy materials and does not concern personal data. Therefore, there are no ethical or legal issues that can have an impact on data sharing and there is no need for informed consent for data sharing as the subjects under study are materials not people.

6. Other issues

STREAMLINE does not use other national/funder/sectorial/departmental procedures for data management.