

Proposals for NORMAN Joint Programme of Activities 2025

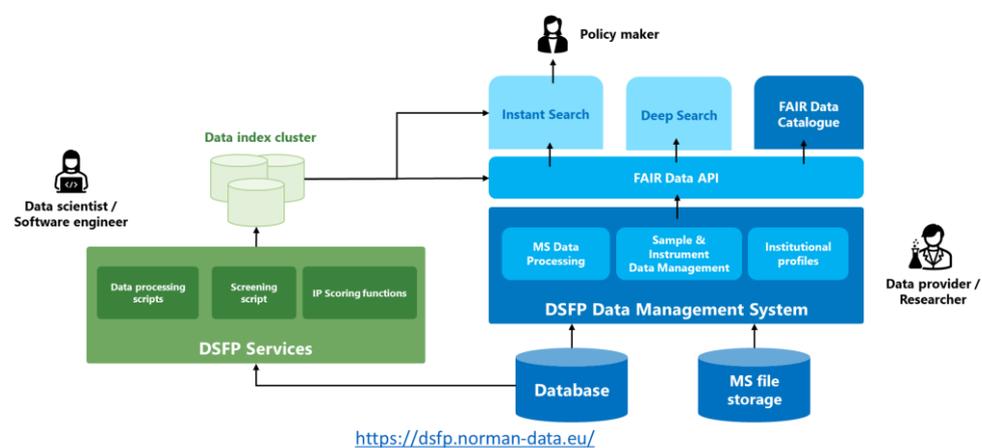
| | |
|---------------------------|--|
| Title | Digital Sample Freezing Platform (DSFP) |
| Type of activity | Research and database development |
| Leader | EI/NKUA (Nikiforos Alygizakis) |
| Topic / activities | <p>Background / Justification for the proposed activity: DSFP developed from a prototype to a production-ready system with enhanced informatic capabilities over the last years. DSFP enables archiving, processing, analysing, data mining and retrieving information for thousands of contaminants of emerging concern (CECs) contained in high-resolution mass spectrometry (HRMS) data. DSFP is ready to support challenging and ambitious goals of NORMAN such as the application of non-target screening prioritization and the automatic retrieval of chemical exposure in early-warning systems for chemicals. The database is a truly unique effort to collect HRMS data from environmental samples that enables for risk assessment of CECs, which is then intended for use by the policy makers and regulators. DSFP, as part of NORMAN Database System (NDS), is a valuable asset for the future activities of NORMAN Association.</p> <p>The informatic architecture that was adopted is presented in Figure 1. Briefly, the technical solution consisted of the following elements:</p> <ol style="list-style-type: none"> 1. DSFP Data Management System (DMS) was designed as a central hub for data contribution and administration. Laboratory data managers, desiring to contribute their chromatograms, register through a user-friendly wizard. Each laboratory data manager belongs to an organization and can register instruments, configure setups, and upload and screen their data under dedicated user panels. 2. FAIR Data API keeps track of the latest data assets contributed to the platform. This API ensures interoperability and enables communication between DSFP and other external sources. Moreover, it provides programmatic access to contributions. The FAIR Data API receives input from the data index cluster, storing screening results in a harmonized format. 3. FAIR Data Catalogue serves as the showcase for FAIR data in the repository, listing collections of data, contributors, data licenses, metadata, and all necessary information. 4. Deep Search is the compound discovery application. It scans through component lists stored in a harmonized and structured way, performing filtering and matching operations. It also scores detections based on the newly developed IP score system. Deep Search may take a few seconds to respond to a request for the investigation of compounds in a collection, with response times depending on the number of matches, the number of samples in a collection, and other factors. 5. Instant search is the application that utilizes fast indexing technologies and NoSQL approaches to display results from Deep Search screenings. The adopted approach is advantageous because it allows a rapid search for compounds in sample collections and enables interconnection with external databases. 6. DSFP services host R and Python functions in Docker images, accessible via RESTful APIs. The output of DSFP services is stored in a data index cluster. <p>More information about the applied technical solution is available in the project proposals of previous years.</p>  <p>https://dsfp.norman-data.eu/</p> |

Figure 1. Schematic of the technical design of DSFP: The Data Management System (DMS) enables laboratory data managers to control and upload samples. The DMS is connected to a SQL database and a file storage system. It interfaces with a FAIR Data API, providing data to the FAIR Data Catalog, Deep Search, and Instant Search. A set of DSFP services, including data processing scripts, IP scoring functions, and semiquantification functions, is deployed in a separate block (DSFP services). These

services allow the execution of functions on the data stored in DMS and MS file storage. Processed results are indexed for rapid data retrieval in Instant Search and the FAIR Data API.

As of September 2024, the platform has been populated with HRMS data of 5,080 unique samples (**Figure 2**) around Europe and beyond. A steady increase in numbers of samples was achieved in 2024. The number of samples is expected to increase exponentially during the next years due to the created facility and the increasing participation of NORMAN Association members in significant projects. The samples covered environmental samples: wastewater (35.77%), biota (28.01%), surface water (25.73%), sediment (4.41%), groundwater (2.74%), soil (1.42%), and other matrices (1.93%).

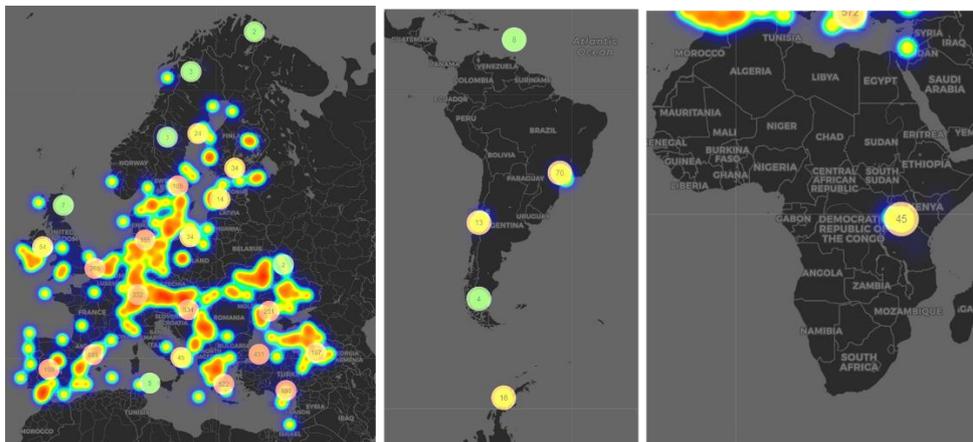


Figure 2. Spatial coverage of samples contributed to DSFP as of September 2024

In 2024, several significant achievements were made. DSFP output elements were harmonized using controlled vocabularies for metadata, component lists, and screening outputs. A metadata schema was defined for all DSFP entities (<https://dsfp.norman-data.eu/data-schema>). The screening process for collections was debugged and optimized, and screening and indexing were applied to five collections. APIs were enhanced, and a new API for instant search was introduced. A web analytics collection system was installed, and submission forms were revised. A map of contributions was added to the main page, and users from various projects received support to ensure effective use of DSFP.

Description of the proposed activity and expected outcomes for 2025:

The continuous development of DSFP's functionalities is crucial for expanding its usage and further enriching the database. Enhancements that improve DSFP's efficiency and encourage researchers to upload their data, thereby increasing the collection of HRMS data, remain a key objective. To address this, the JPA proposal focuses on maintaining and advancing DSFP's functionalities:

1. Preparation of a manuscript describing the new technologies utilized in upscaling DSFP
2. Screening and indexing all collections in DSFP to enable instant search capabilities and populate EMPODAT-SUSPECT.
3. Developing a functionality for batch import of new samples.
4. Further integration with DataCite (assigning DOI to each contributed dataset)
5. Creation of a machine learning plugin for analyzing and further exploiting NTS data.
6. Exploration of collaboration opportunities with other major platforms (e.g., MassBank consortium, GNPS).
7. Improving guidance documents and producing instructional videos for DSFP.
8. Maintaining and supporting current and future users, including fostering connections with PARC and other EU-funded projects involving NORMAN members.

Added value / Link with other NORMAN activities and / or other projects

- Cross-Working Group Activity on Non-target Screening (NTS)
- More case-studies and support for prioritization exercises of NORMAN WG1 Prioritisation
- Added value for the NORMAN Database System
- Synergy with WP4 and WP8 of the PARC project
- A FAIR data management plan solution for NORMAN laboratories for the EU-funded projects
- DSFP as the place for hosting HRMS data of NORMAN collaborative trials

| | |
|--|---|
| Participants | EI, NKUA, NILU, LCSB, Eawag, UFZ and all NTS members of NORMAN. |
| Proposed in-kind contribution | Working hours for implementation the project. |
| Contribution needed from NORMAN Association | Total funds required for IT support for the application of the suggested improvements: 15,500 € |