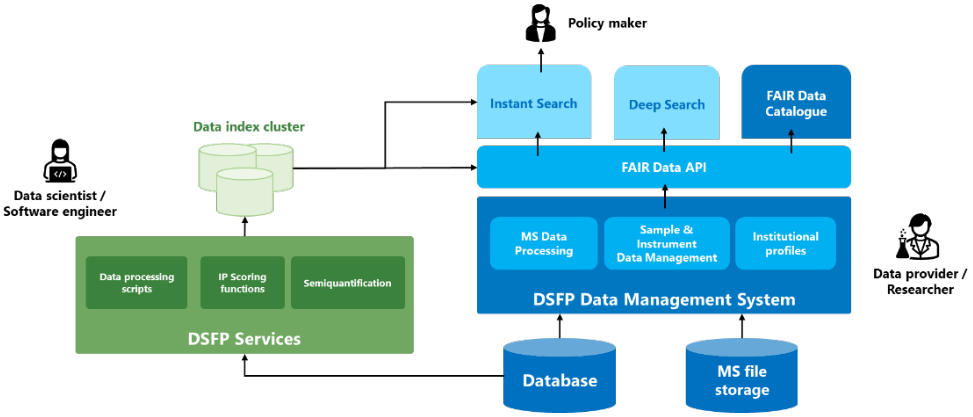
Proposals for NORMAN Joint Programme of Activities 2024

Title	Improvement of the DSFP
Type of activity	Research and database development
Leader	Nikiforos Alygizakis (EI/NKUA)
Topic / activities	***Background / Justification for the proposed activity:*** The previous activities on upscale and improvement of Digital Sample Freezing Platform (DSFP) were implemented successfully establishing a significant infrastructure for NORMAN Association. The purpose of the activities was to bring the prototype to a production-ready facility with improved informatic characteristics (usability, maintainability, scalability, performance and extensibility). The activities allowed the implementation of FAIR principles and the establishment of a well-document API for the automatic retrieval of the occurrence of contaminants in stored collections. The technical design of DSFP was refined and is presented in **Figure 1**.  Policy maker Instant Search Deep Search FAIR Data Catalogue Data index cluster FAIR Data API Data scientist / Software engineer MS Data Processing Sample & Instrument Data Management Institutional profiles DSFP Data Management System Data provider / Researcher Data processing scripts IP Scoring functions Semiquantification DSFP Services Database MS file storage **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. Briefly, the technical solution consisted of the following elements: 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. More information about the applied technical solution is available in the project proposals of previous years. DSFP is in position to facilitate fast publishing, effective management and open access of research assets. The platform enables archiving, processing, analysing, data mining and retrieving information for thousands of contaminants of emerging concern (CECs) contained in high resolution mass spectral (HRMS) data. 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.

As of September 2023, the platform has been populated with HRMS data of 3,870 unique samples (**Figure 2**) around Europe and beyond. Despite the ongoing infrastructure developments, a steady increase in numbers of samples was achieved in 2023. 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 such as the PARC project. The samples covered environmental samples: surface water (28.68 %), biota (22.97%), wastewater (24.21%), sediment (3.41%), groundwater (3.49%), soil (1.86%), and other matrices (2.82%) and human biomonitoring (12.56%).

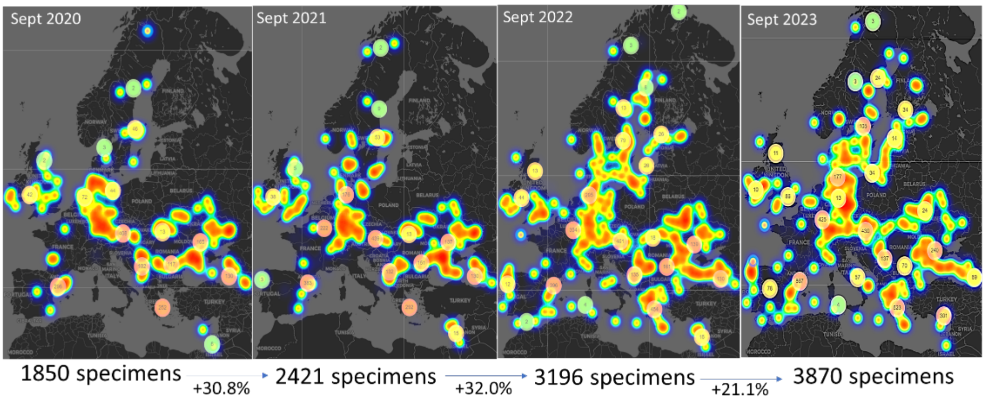


Figure 2. Spatial coverage of samples contributed to DSFP as of September 2023

Description of the proposed activity and expected outcomes for 2024:
DSFP has been used in a series of sampling campaigns and collaborative trials of NORMAN Association, providing satisfactory results in terms of identifying substances usually overlooked by target and non-target screening. DSFP is nowadays frequently used in various large-scale monitoring campaigns and acting as a safety-net for the detection of potentially hazardous substances. DSFP has proven to be useful for various activities of the NORMAN network, especially at the prioritisation of CECs and as an early-warning system for chemical risks. DSFP, as part of NORMAN Database System (NDS), is a **valuable asset** for the future activities of NORMAN Association. The continuous development of functionalities of DSFP is of importance to expand its use and further enrich the database. Changes that would make DSFP more efficient and that would stimulate the interest of researchers to upload their data and thus increase the collection of the HRMS data remain still a very important objective.

For this reason, the JPA proposal aims to maintain and improve the functionalities of DSFP:

1. Screening and indexing all collections in DSFP to prepare them for EMPODAT-SUSPECT.
2. Incorporating a statistical plugin for the analysis of NTS data (unsupervised methods).
3. Applying mixture evaluation using network analysis.
4. Making the standardized NTS workflow available to the scientific community using Docker technology. This approach allows the execution of the NTS workflow at a high scale in an isolated virtual environment containing all necessary software and dependencies.
5. Standardizing the output elements of DSFP using controlled vocabulary for metadata, component lists, and screening outputs.
6. Creating API mechanisms for predicting all necessary information for importing new compounds.
7. Testing 4D-HRMS data from all vendors and updating the DSFP cookbook.
8. Investigating the possibility of integrating DSFP with MassBank and RMassBank.
9. Creating guidance documents and videos for DSFP.
10. Preparing a manuscript describing the new technologies utilized to upscale DSFP.
11. Introducing necessary functionalities to allow the use of DSFP in other fields, such as human biomonitoring.
12. Maintaining and supporting current and future users, including connection with PARC and other EU-funded projects of 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: 14,000 €