# Proposals for NORMAN Joint Programme of Activities 2025

| | |
|---|---|
| **Title** | **NORMAN Database System (NDS)** |
| **Type of activity** | Database maintenance and continuous update |
| **Leader** | EI |
| **Topic / activities** | ***Background / Justification for the proposed activity:***<br>The NORMAN Database System (NDS) is a joint activity of all NORMAN members and at the core of the NORMAN activities, providing data and tools to fulfil its goals and visions. The NDS consists of 13 integrated databases modules:<br>1. Suspect List Exchange - https://www.norman-network.com/nds/SLE/<br>2. Substance Database - https://www.norman-network.com/nds/susdat/<br>3. Chemical Occurrence Data (EMPODAT) - https://www.norman-network.com/nds/empodat/<br>4. Ecotoxicology Database - https://www.norman-network.com/nds/ecotox/<br>5. Digital Sample Freezing Platform (DSFP) - https://norman-data.net/Verification/<br>6. Substance Factsheets - https://www.norman-network.com/nds/factsheets/<br>7. NORMAN MassBank - https://massbank.eu//MassBank/<br>8. Passive Sampling - https://www.norman-network.com/nds/passive/<br>9. Antibiotic Resistance Bacteria/Genes - https://www.norman-network.com/nds/bacteria/<br>10. SARS-CoV-2 in sewage - https://www.norman-network.com/nds/sars_cov_2/<br>11. Bioassays Monitoring Data - https://www.norman-network.com/nds/bioassays/<br>12. Indoor Environment - https://www.norman-network.com/nds/indoor/<br>13. Prioritisation - https://www.norman-network.com/nds/prioritisation/<br>All NDS modules can be searched either individually or starting from the module ‘Search All Databases’ (https://www.norman-network.com/nds/common/), where any substance from SusDat can be searched and displayed with all available data for this substance in any of the database modules.<br>The prototype of new module, EMPODAT-SUSPECT (see JPA 2024 WG1), has been tested in 2024 with a test-set of 58 million suspect screening data. It is expected that the full version will be ready in 2025 and gradually filled out with suspect screening data from HRMS chromatograms archived in DSFP (>5,000 samples as of December 2024; screening of >95,000 substances in each sample; assigning IP score to all detected substances (see JPA 2024 DSFP); total >475 million data entries).<br>Another new module NORMAN BIOACTIVITY Database (see separate proposal in JPA 2024; https://www.norman-network.com/nds/badb), has been developed in 2024 and filled out with the first 1,037 entries. The request to test the functionality and performance of the module and newly developed Data Collection Template to contribute to the database has been distributed to all NORMAN members. A launch of the final version of the database is expected in 2025.<br>An automated prioritisation module in the NDS is available for the target substances archived in EMPODAT. There is an on-going effort to develop the prototype of the prioritisation module for suspect substances archived in EMPODAT-SUSPECT in line with the methodology as described in Dulio et al., 2024, DOI: 10.1186/s12302-024-00936-3 (see also JPA 2024 WG-1). A dedicated automated prioritisation module for biota samples has been developed in 2022-2024 allowing for a more refined search using biota-specific queries. It is expected to bring all of these tools together in 2025 and support prioritisation activities of all WGs, with a specific focus on WG-8 (marine environment), Joint Danube Survey 5 (freshwater environment) and WG-7 (terrestrial environment). A set of tools allowing for visualisation of prioritised data have been developed (see presentation of WG-1 and NORMAN GA meetings). Further improvement of the tools and proposal/selection of proper indicators to be tested by NORMAN WGs and CWGs is expected in 2025.<br>Based on the outcomes of the NORMAN Database Workshop in 2023, an API portal has been developed allowing for automated data sharing with external databases (https://www.norman-network.com/nds/api/). It is now available for: NORMAN SusDat, EMPODAT, Ecotoxicology Database and Passive Sampling Database. All modules will contain API including documentation (in progress). In order to enhance the FAIRness of NDS, on-line accounts for assigning DOI to all contributed datasets has been developed for EMPODAT and tested using https://doi.datacite.org/. Accounts created for each NDS module and contributed dataset will continue to be developed in 2025.<br>All NDS modules underwent continuous update in 2024. The EMPODAT database was enlarged for more than 600,000 new data entries and contained 96,625,121 data in the end of 2024; with monitoring data for 4,687 substances. A close cooperation with PARC, PROMISCES, SPRINT, CONNECT 2 and EU4EMBLAS projects have been established resulting in inflow of valuable datasets in harmonised formats. Several datasets are still not publicly accessible until finalisation of the projects and related reports/publications. A significant effort has been put into implementation of the automated protocol requested by the EC JRC for transfer of EMPODAT into IPCHEM on an annual basis.<br>In the NORMAN Ecotoxicology Database, new EQS and PNEC values were uploaded and available PNECs underwent continuous revision with a special effort to remove duplicate entries. DCT for PNECs |

from studies/literature was updated and all NORMAN members are invited to contribute, especially with the experimental values. Based on the specific request of the WG-8 (marine mammals) and WG-7 (terrestrial environment), rat toxicity threshold values predictions were produced for SusDat substances using VEGA QSAR model. The values are ready to be integrated into both the database and automated prioritisation tool in 2025. Several models have been identified within the network as suitable candidates to provide reliable predictions for (eco)toxicity threshold values. There is a need to decide how to incorporate these values for numerous toxicity endpoints (e.g. ToxAI aiming at 105 different endpoints) into the Ecotoxicology database, which is currently based on aquatic ecotoxicity-derived threshold using the three 'regulatory' trophic levels (algae, daphnia, fish). At present, the marine and terrestrial endpoints are of particular interest to NORMAN members.

In the NORMAN Antibiotic Resistant Bacteria/Genes Database have been developed five new DCTs for ARG and ARB matrices (water, soil, sewage sludge, plant crop, air). New data were added and the progress was reported in Alygizakis et al., 2024, DOI: https://doi.org/10.1016/j.watres.2024.121689.

The NORMAN Indoor Environment Database has been updated and upgraded DCT is available for providing new data ( https://www.norman-network.com/nds/indoor/downloadDCT.php) in 2025. For an overview, see Haglund at al., 2024, DOI: 10.1016/j.scitotenv.2024.177639.

The NORMAN Passive Sampling Database contained 4,213 target analyses data entries in the end of 2024. However, many of the efforts of the CWG-PS aim at the collection of both target and suspect screening data (e.g. Joint Danube Survey 5 in 2025). There is a need to create a database to host such data.

The SLE contained in the end of 2024 120 lists (cf. separate JPA 2025). There is a need to add substances from the newly added lists in 2024 into SusDat, which contained 120,916 substances in December 2024. Additionally, it is necessary to collect all supporting data required for suspect screening (cf. separate JPA on DSFP) and prioritisation (cf. separate JPA WG-1) of the new substances.

In SusDat, the automated curation tools supported with the manual control resulted in removal of all duplicate entries based on the name or CAS No. of the substances. There are however 253 duplicates based on the StdINChIKey, which still need attention and appropriate strategy how to deal with them. The 'batch conversion of identifiers' tool has been refined (see https://www.norman-network.com/nds/susdat/susdatConversion.php) allowing for seamless conversion among names / CAS. Nos./ StdINChIKey/ NORMAN SusDat ID of any (group of) substances in SusDat. All contributors to the NDS in general, and SLE in particular, have been encouraged to assign NORMAN SusDat ID to the contributed substances in any of the NDS modules. Once NORMAN SusDat IDs are generated, they can be directly used for batch searches, e.g., in EMPODAT, DSFP or in the prioritisation modules.

Substance Factsheets were being updated for latest information from interrelated databases in 3-months interval in 2024. In order to support the hazard assessment, development of a new module "Hazards and Properties" has been proposed, which will generate data for both the "NORMAN Substance Factsheets" and "Prioritisation" modules in 2025. This is an attempt to support the EU Chemicals Strategy and the 'One Substance One Assessment' platform for (i) physico-chemical properties, (ii) fate and transport, (iii) PBMT derivation, (iv) PBMT classification, (v) CMR and ED classification; all supported with appropriate Data Collection Templates. An initial document has been developed and discussed at the WG-1 meeting in November 2024. Data on PBT, CMR and ED properties have already been generated for more than 76,000 SusDat substances using the JANUS tool (VEGA platform). Their upload into the prototype of the new module is scheduled for 2025.

The NORMAN Artificial Intelligence Workshop took place in Leipzig in October 2024. The discussion addressed also various aspects of future development of the NDS. Some short-term proposals were to (i) develop a list all open access models used to feed various NDS modules and (ii) archive all training and validation datasets of these models. In the follow up discussion, models used in the TerraChem project (terrestrial environment and biodiversity) and ToxAI tool (toxicity thresholds for numerous endpoints) were proposed as a starting point. All members of the network are invited to contribute. The long-term vision is to establish a transparent link to all source data/parameters used for model predictions (including those not yet stored in the NDS) as a part of the EU Common Data Platform and One Substance One Assessment strategy.

Many of the NORMAN members are using various text mining tools to obtain data from peer-reviewed or grey literature for various purposes, e.g. for establishing early-warning systems for chemicals in the environment. However, the experience shows that these data are in various formats difficult to compare, often aggregated and without sufficient data quality information. There was a proposal to establish a harmonised data format and European repository for such data within the NDS. Already collected data on metals and organic pollutants in apex predators from the TerrraChem project were suggested as a starting point. An attempt to incorporate soil pollution data from the EC LUCAS project (aggregated data, no coordinates) was proposed.

Automated curation of spatial data in EMPODAT was presented at the workshop using CleanGeoStreamR script developed at UFZ Leipzig (open access at GitHub). It has been agreed to process the latest version of EMPODAT (>96 million fdata entries) with the tool to correct wrongly assigned coordinates, misspellings

| | |
|---|---|
| | of station/river/sea region names etc. in the originally provided data. In the test run, more than 107 million corrections were made. It is planned for 2025 to incorporate these corrections into EMPODAT and develop a strategy for flagging the corrected data. Ultimately, the script could be given to data providers for automated raw data treatment prior to sending it to the NDS.<br>***Description of the proposed activity and expected outcomes for 2025 (and beyond):***<br>The task for maintenance of the NDS and its continuous upgrade in 2025 include:<br>• Development of the 'Hazards and Properties' database module (in collaboration with WG-1);<br>• Further development of API portal allowing for automated data sharing with external databases.<br>• Assigning DOIs to all contributed datasets in the NDS.<br>• Further interlinking of all NDS modules and quality check of all input data;<br>• NDS Chemical Occurrence Data (EMPODAT): maintenance, upgrading and feeding of new data into the database; sharing the data with IPCHEM;<br>• Development of a database module for archiving data obtained by data mining tools from publications, patents and grey literature; establishment of a workflow for their processing into the 'NORMAN format' (as a possible joint collaboration with TerraChem project)<br>• Further development/upgrade of visualisation tools in the NDS;<br>• Continuous upgrade and maintenance of SusDat;<br>• Update of Passive Sampling module with new datasets; programming of Passive Sampling – SUSPECT module;<br>• Upload of new data into the ARBs/ARGs module (budget already covered under WG-5);<br>• Upgrade of Substance Factsheets module – systematic collection of all data needed for prioritisation and data download functions;<br>• Development of the new module for collection of training and validation datasets for models feeding various parts of the prioritisation framework of NORMAN (cf. JPA 2025 proposal for Blueprint for linking ecotoxicity to different levels of biodiversity damage);<br>• Correction of spatial data in EMPODAT based on the curation by CleanGeoStreamR script and development of strategy for flagging correction of original data (collaboration with UFZ).<br>***Added value / Link with other NORMAN activities and / or other projects***<br>The proposed tasks will benefit all WGs and CWGAs in the NORMAN network. |
| **Participants** | EI, all interested members |
| **Proposed in-kind contribution** | All – contribution of existing data<br>EI – overall coordination |
| **Contribution needed from NORMAN Association[1]** | Maintenance and continuous update of the NDS:<br>- EI: 42,000 €<br>Rental of the server hosting the NDS, management and backup system:<br>- EI: 8,600 € |

[1] Please, provide here a transparent justification of the requested resources and of the in-kind contribution, thereby distinguishing between the costs associated with "person-months" for the organisation, the "travelling costs" for invited speakers and the costs for the logistics (e.g. meals, room rental etc.)