Data Management

NanoReg2 presents experience with integrating large sets of nanosafety data generated from past NanoSafety Cluster projects with the help of a substance data model (Fig 1), implemented by the eNanoMapper database software This data model is also successfully used to handle chemical substances and safety data from ECHA dossiers.

While multiple structured import formats are supported (IUCLID, RDF, JSON), the nanosafety data from past and ongoing projects use custom spreadsheet templates, currently encompassing over 1000 Excel files. Import of Excel files is enabled by a configurable parser that maps the spreadsheet data via external configuration files. Multiple export formats are supported, including tab delimited files, RDF and ISA-JSON.
The eNanoMapper ontology is used for harmonisation of the terminology and as a synonym list for query expansion.
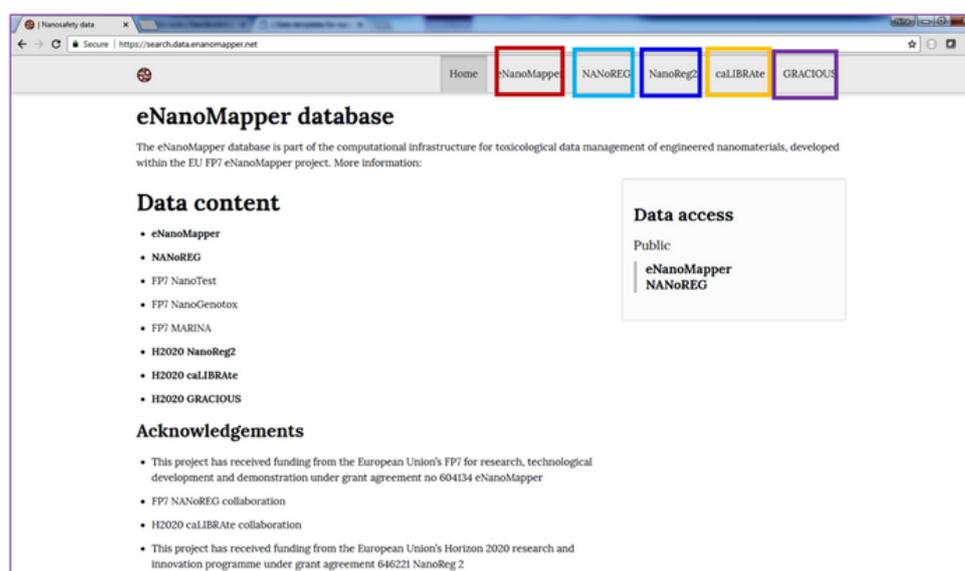

Fig 1. The data integration relies on a common data model and identifiers mapping using the eNanoMapper ontology

Data generated by NANoREG, NanoTest, MARINA, NanoGenotox is compiled, annotated and imported into eNanoMapper database instances. The database architecture follows the decision proposed and agreed during NanoReg2 Data Solution team meetings is illustrated on Figure 2. These databases offer a user friendly web interface and REST API and serve as building blocks to provide federated search across all or subsets of the database instances, enabled by Apache Solr backend.

The setup is composed of describes multiple installations (instances) of eNanoMapper databases, populated with specific projects data (Fig 2, bottom). Project-specific search applications, provide integrated view of selected sets of database instances, without copying the data (Fig 2, top). The web site at https://search.data.enanomapper.net serves as an point to all integrated search views. Restricted access for project partners was introduced. Improvements in search interface, export data format and summaries based on partners feedback. Most popular data access is download query results as text files. REST API usage is reported only by two partners.

Data sources: 1) The NANoREG TNO database content is converted into eNanoMapper database SQL and updated several times 2) Excel spreadsheets, roughly following either JRC "ISA-Tab-Nano logic" templates, or IOM dose-response templates. The common enanoMapper data model enables translation between multiple input and output formats.

Data cleanup: 1) harmonizing identifiers of materials, cells, methods, etc. 2) adding missing data identified by partners; fixing data errors; 3) Adding derived data (e.g. IC20) 4) introduced linking between measurements from different assays (e.g. experiments made in parallel in order to characterize the NMs and toxicity assessment) 5) Exposure data (provided by TNO as exported from NECID) was recently added to NanoReg2 DB.

Challenges for the purpose of grouping:
Data interpretation is difficult due to sparsity of the data and the lack of consistency between different data sources, SOP, publications, data generation span ~ 10 years.

Scoring for the purpose of better modelling is suggested by NanoReg2 WP 1 as a solution to the data interpretation need. Grouping is largely dependent on defined parameters and thresholds of toxicity (i.e. IC10, IC20 or other parameters describing e.g. dose-response). Methods for harmonized interpretation of data are needed for this purpose.
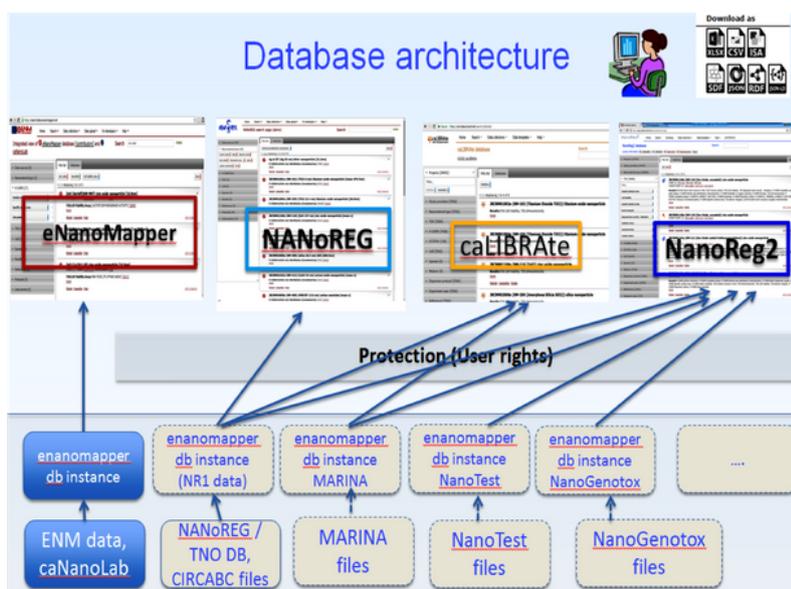

Fig 2. Free text and faceted search applications, with public and restricted access for different subsets of data, are available at https://search.data.enanomapper.net

Results and Conclusions

- The NanoReg2 integrated database allows project partners to access data from past EU FP7 funded projects through a common view and faceted search. The database is actively used by project partners, helping to identify and, where possible, resolve data quality and completeness issues. The same technology and content is used by several H2020 projects (caLIBRAte, PATROLS. Gracious) and the EUON - European Union Observatory for Nanomaterials.
- Scoring (and scaling) approaches that allow for different data (including different assay-specific metrics, as well as results from diverse assays) to be integrated into a final score, indicative of a specific toxicity endpoint, will be useful for enhanced modelling towards establishing grouping hypotheses.