From heterogeneous, multi-source data to harmonized datasets: a major challenge for the assessment of soil functions. The LANDMARK project approach

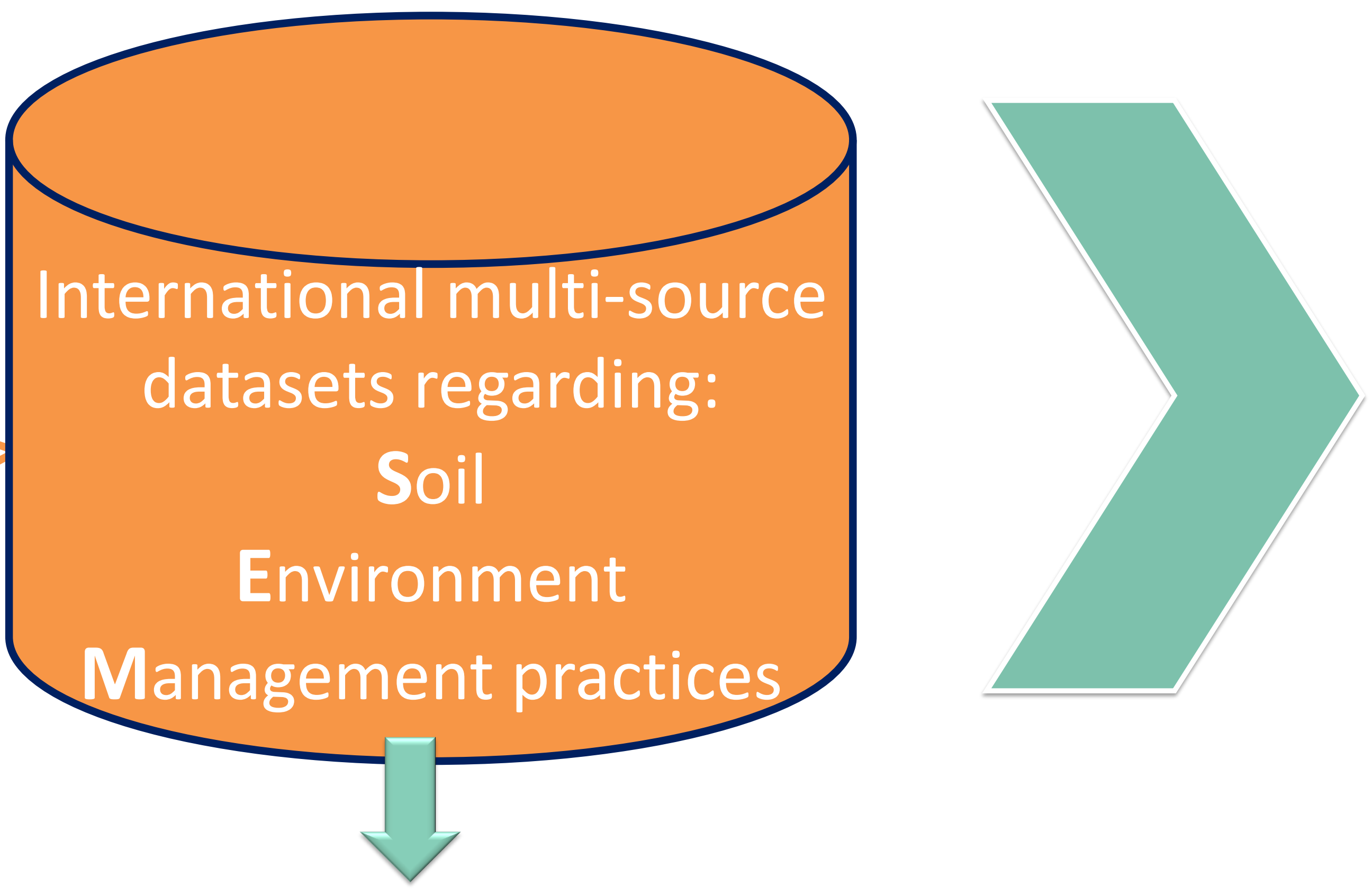
Sébastien Drufin, Nicolas Saby, Marion Bardy, Erika Micheli, Benoît Toutain,
Rachel Creamer

From heterogeneous, multi-source data to harmonized datasets: a major challenge for the assessment of soil functions.

S.Drufin[1], N. Saby[1], M. Bardy[1], E. Micheli[2], B. Toutain[1], R. Creamer[3]
[1] INRA, unité InfoSol, US 1106, F – 45075 Orléans Cedex 2, France
sebastien.drufin@inra.fr, nicolas.saby@inra.fr, marion.bardy@inra.fr, benoit.toutain@inra.fr
[2] Szent István Egyetem • 2100 Gödöllő, Páter Károly utca 1
micheli.erika@mkk.szie.hu
[3]Department of Soil Quality, Wageningen University, Droevendaalsesteeg 4, NL-6708 PB Wageningen, P.O. Box 47 , The Netherlands
rachel.creamer@wur.nl

Background & objectives

LANDMARK[1] is a H2020 pan-European multi-actor consortium of leading academic and applied research institutes, chambers of agriculture and policy makers that will develop a coherent framework for soil management aimed at sustainable food production across Europe. The LANDMARK proposal builds on the concept that soils are finite resources that provide a range of ecosystem services known as "soil functions". Functions relating to agriculture include: primary productivity, water regulation & purification, carbon-sequestration & regulation, habitat for biodiversity and nutrient provision & cycling.



- ✓ **Develop** a comprehensive scientific "SxExM" framework for understanding and quantifying **soil functions** in relation to soil properties (S), climatic zone (E) and soil management (M), stratified by land use.

Challenge = gather and harmonize multi-source data in a EU context

Material & Methods

- Determination of a proper schema for data and a framework for all the different transformations applied to data. Multi-sources data origin calls for the use of suitable tools of Business Intelligence Information System. Data are collected according to the needs expressed by research teams in charge of data processing.

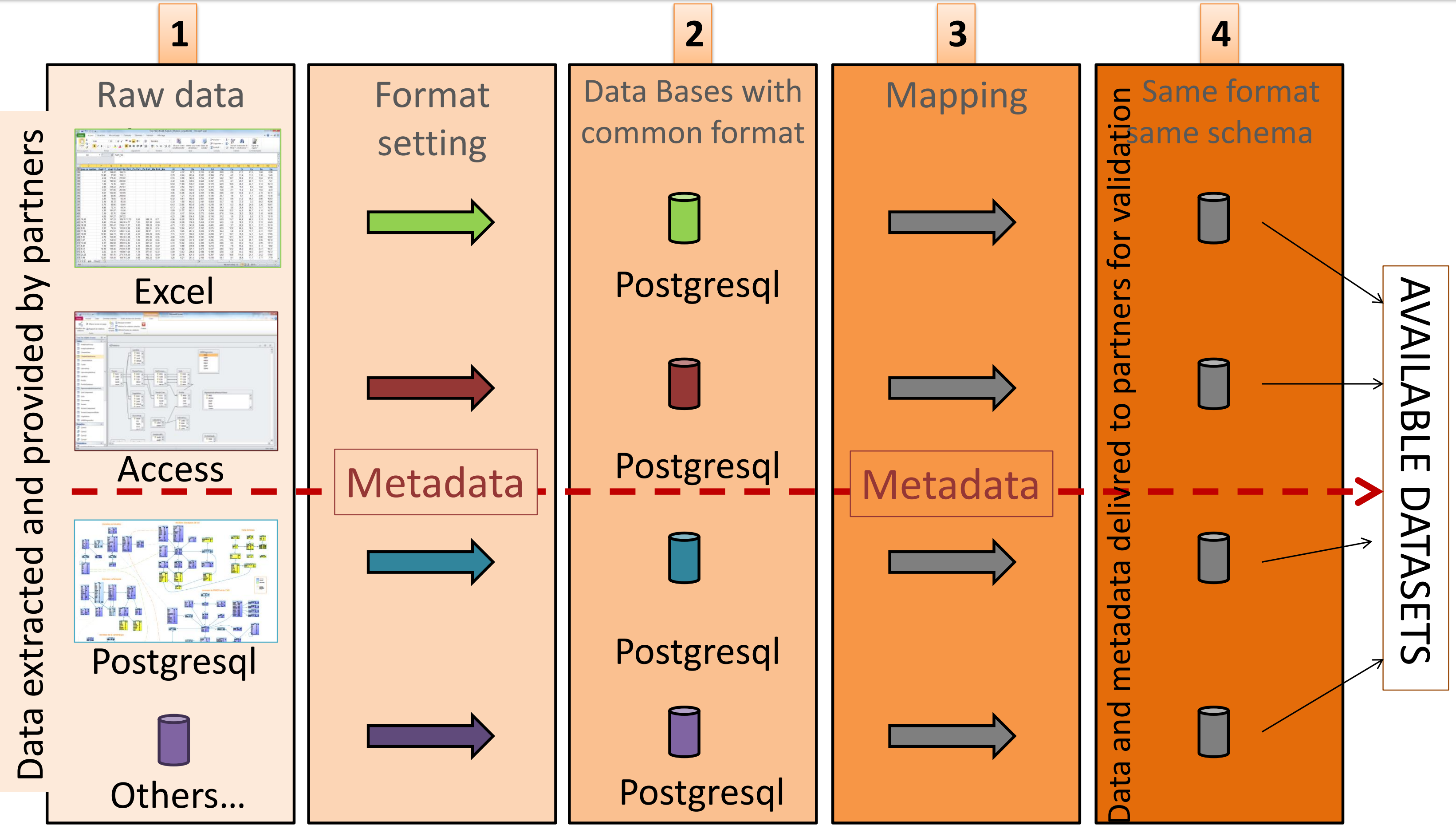
- Set up of the computer infrastructure: creation of a data warehouse Transformation of dataset from "raw" data to datamarts, according to the different steps :

1 - Data are collected in various formats (operational systems);

2 - Data are transformed in order to match with database system (data staging);

3 - Data are then related to reference data in order to relate then to common typologies, etc. (data warehouse) and thus be harmonized, using Extract, Transform and Load (ETL) processes;

4 - At last, harmonized data are gathered in datasets (called datamarts) and delivered as outputs for the algorithms and models that will compile data. The LANDMARK project will develop harmonized datasets for the assessment of each of the soil functions.



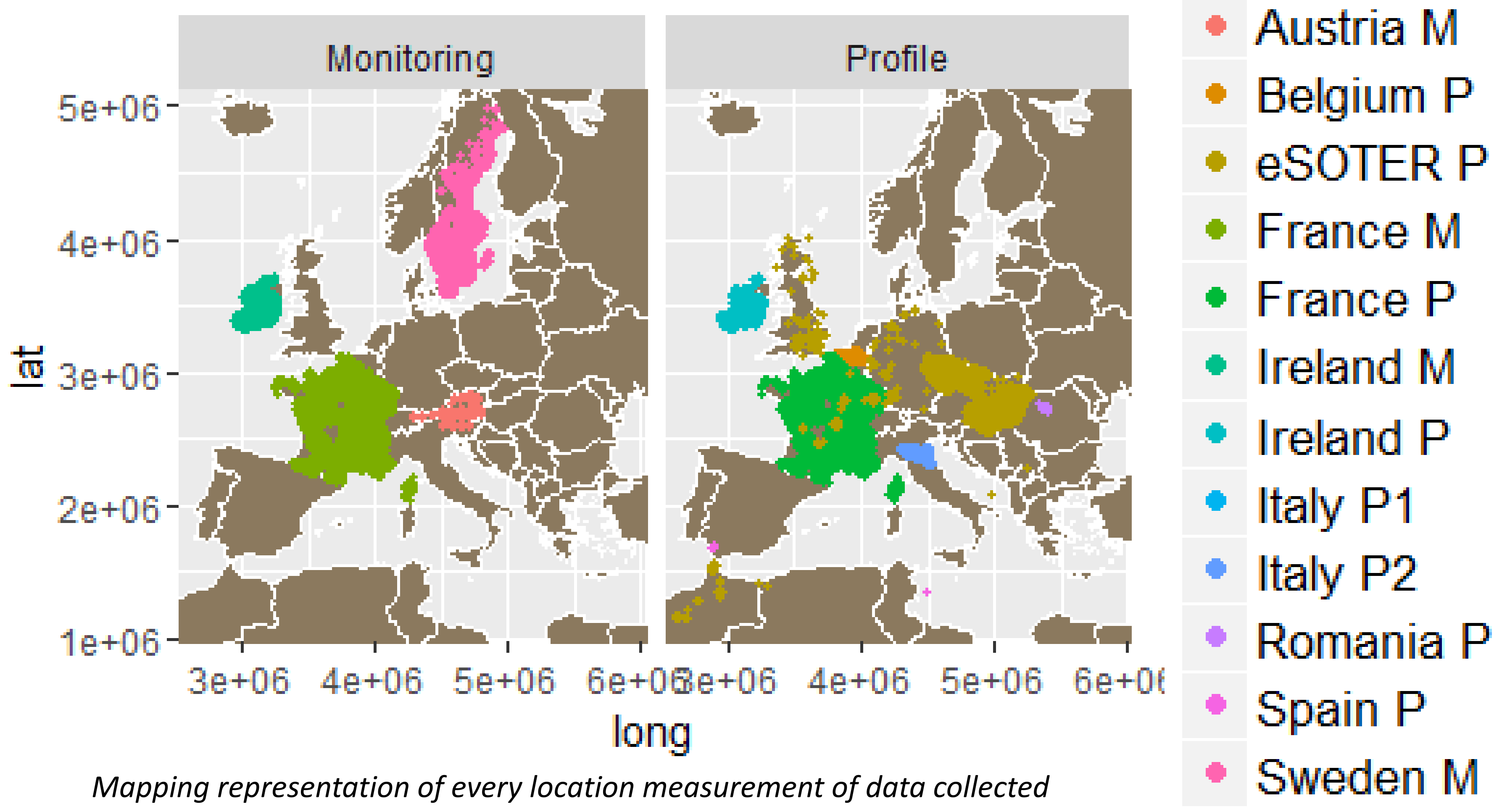
Data integration and harmonization process from heterogeneous raw data providing to homogenous datamarts delivered for soil function modeling.

Intermediate Results

Soil attributes are the first type of data collected from 11 datasets yet. Via ETL tools, from those datasets, 29 different physical, chemical, descriptive attributes have been integrated within the data warehouse. An amount of **988.494** workable data is now available for processing.

Data type	Country	Profiles	Horizons
Profile data	Belgium	6889	41789
	Spain	47	196
	France	1364	4353
	Italy	5612	46481
	Romania	10	30
	Hungary - Czech republic	2213	9077
	Ireland	806	3628
Monitoring program	Ireland - Topsoil	1310	-
	Sweden - Topsoil	5182	-
	France - Composite	2215	4243
	Austria - Composite	139	973
Total location measurement		25787	110770

Table summarizing the amount of location measurement available per dataset provided.



Mapping representation of every location measurement of data collected according on whether data are from monitoring program or profile descriptions.

What's next ?

- Metadata description is in progress, being written directly into tables in the data warehouse as comment for each fields, with a track of every modification done on each raw data during the harmonization – mapping step via ETL tools.

- On these data, a diagnostic horizon computation will be done. Scripts that will be used to proceed those calculations will be implement into ETL treatment to produce estimations of diagnostic horizon that will be directly integrated as a complementary attribute in the soil.

Only soil data have been integrated. In the next step, environment and land management data will be added into a new table within the data warehouse.

European data (e.g. Corine land cover, CAPRI, etc…) will also be integrated within the data warehouse to inform soil profiles on land management.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 635201

LANDMARK

www.landmark2020.eu email landmark@teagasc.ie twitter @Landmark202