Work package number 1	Lead beneficiary CSIC				
Work package title	**Data knowledge discovery**				
Participant number	2	2	8	5	3
Short name of participant	EBD-CSIC	IFISC-CSIC	URV	EPFL	SCITE
Person month per participant	24	24	6	6	6
Start month	1				
End month	24				

Objectives

- To develop an evolutionary biology-inspired semantic framework for data discovery
- To derive semantic functionality rules required for data computation discovery
- To apply data discovery properties for the Oceans' sustainability case study

Description of work

Task T1.1: Evolutionary semantic algorithms (ESA) (M1-M18) *Leader:* ***EBD-CSIC****. Contributors: 2*

ESA will find classes and datatype properties from ontologies, and raw data from non-semantic databases. ESA will infer semantics on the raw data to link them to the ontological terms. We will translate the semantically-annotated databases to a $Neo4j$ graph database by mapping classes to nodes, object properties to links between nodes, and datatype properties to nodes' attributes. The graph database has an architecture flexible enough to get high scalability to accommodate many source data and to infer its properties using multilayer metrics (T1.2). T1.1 provides ESA to allow WP2 and WP3 to implement the models for causal knowledge discovery and discovery in federated networks.

Task T1.2: This task extends T1.1 into multilayer network metrics for general principles of data discovery (M1-M18) *Leader:* ***IFISC-CSIC****. Contributors: 2*

Multilayer network metrics for ESA will focus on data heterogeneity to explore how data configurations, privacy requirements, formats, dimensions, biases and spatiotemporal resolution affect data discovery properties [24–26].

Task T1.3: Based on the framework developed in T1.1 and T1.2, URV will derive automation procedures for data knowledge discovery (M15-M21) *Leader:* ***URV****. Contributors: 8*

Automation rules identify the ESA rules for data discovery [16]. URV will complement T1.1 and T1.2 to obtain posterior probabilities of evolutionary expressions that represent the empirical patterns of the data knowledge graph generated in T1.1 and T1.2.

Task T1.4: Reproducible data knowledge graphs (M15-M21) *Leader:* ***EPFL****. Contributors: 5*

In this task the EPFL will integrate the work done in T1.1 and T1.2 into reproducible and replicable data knowledge graphs. T1.4 samples the data sources to obtain the robustness of data heterogeneity. Robustness will be analyzed working closely to the IFISC-CSIC partner in T1.2.

Task T1.5: Visualize (M15-M21) *Leader:* ***SCITE****. Contributors: 3*

In this task the partner SCITE will apply visualization algorithms to the patterns obtained in T1.1 and T1.2. Data knowledge graphs will be represented in static (figures) and dynamic (animations) visualizations using cutting-edge graphic libraries like D3.js, LightGraphs.jl. All animations will be used by SCITE to strengthen the dissemination, communication and exploitation activities.

Task T1.6: All participants apply results from ESA and multilayer network metrics into a fully automated, reproducible and animated Oceans' sustainability case study (M15-M24) *Leader:* ***EBD-CSIC****. Contributors: 2,3,5,7,8*

ESA and multilayer network metrics will generate the sustainability of the Oceans data knowledge graph integrating many data sources. Fishery data (i.e., global fishing watch), species interactions data, environmental data and social and stakeholders groups data with different interests within each country, etc, will be merged into the sustainability of the Oceans database started in 1965 containing around 9 million entries, 1612 species, 20 countries and 11 sampling methods (Figure 2).

Deliverables

D1.1 Semantic evolving software for data discovery **(M18)**
D1.2 Report on definition of multilayer network metrics applied to data discovery **(M18)**
D1.3 Automated demonstrator of evolutionary semantic rules for data discovery **(M21)**
D1.4 Reproducible demonstrator of evolutionary semantic rules for data discovery **(M21)**
D1.5 Visualization demonstrator of evolutionary semantic rules for data discovery **(M21)**
D1.6 Demonstrator all parts for the Oceans' sustainability case study **(M24)**