

ICIS Meeting summary

Venue

FAO-Rome, 26-28 January 2009

Participants:

S. Tsuji, M. Taconet, F. Calderini, C. Baldassarre, A. Gentile, A. Ellenbroek (FAO Rome)
Pasquale Pagano (Consiglio Nazionale delle Ricerche CNR - Italy)
Leonardo Candela (Consiglio Nazionale delle Ricerche CNR - Italy)
Donatella Castelli (Consiglio Nazionale delle Ricerche CNR - Italy)
George Kakaletris (University of Athens - Greece)
Pavlos Polydoros (University of Athens - Greece)
Diego Milano (University of Basel - Switzerland)

Functional analysis result

The aim of the meeting was to establish the goals, scope and a tentative implementation plan to support an Integrated Capture Information System (ICIS) on the D4Science infrastructure.

The main function of the system will be to provide a quality repository for ICIS data. It shall achieve this by establishing facilities for data import, harmonization, re-allocation, retrieval and export across different reference schemes and data-formats.

The main results at technical implementation level are to:

- Incorporate statistical time series and reference data in the D4S infrastructure under a relational database format,
- Provide environment enabling to import, search, manipulate and retrieve data,
- Integrate effectively main functionalities of the FishStatJ desktop application,
- Offer FAO and Regional Reference data in all data-mapping scenarios, i.e. allow users to map data-sets not only using one standardized mapping schema, but also using other established mapping schemes.
- Ensure that clear references be kept to the origins of data and batches,
- Make clear distinction between original source data and modified/processed data and ensure that modified/processed data not be taken erroneously as original source data.

The developers also discussed an implementation plan and work-schedule, but these will be re-adjusted with upcoming specification development and workload assessments.

The development will focus on delivering a working prototype before the summer; hence several strategic decisions were made to be able to meet this dead-line:

1. Import data can only take a preformatted format.
2. All data will be database-administered.
3. Output will be table-data centric.

After the prototype has been delivered, these decisions can be revised.

Meeting results

1. In the first phase, data import will be enabled for normalized data in comma separated values (csv) files. These will initially be prepared by FAO (A. Ellenbroek). The D4Science team can thus focus on the core of the system, i.e. the database and application layers that are new to the infrastructure.
2. D4Science will enable the import of statistical time series and reference data into the infrastructure. Data will be mapped using db-stored relationships. The incorporation of ontology-driven mapping rules is not foreseen in the prototype phase.
3. All imported data will be visible in the new infrastructure, as well as the status of the import and success. The infrastructure will inform the user of import success at the file level (e.g. report the number of lines imported).
4. The D4S infrastructure will enable the harmonization of heterogeneous and dispersed datasets into a common repository, format, model, and structure. Harmonized data keep their original values, but are already available for common services (search, browse, query, and export).
5. After harmonization, data can be compared to other data-sets, for which mapping rules are required. These mapping rules may be derived from diverse classifications, but can be repeated across different classifications. E.g. a dataset can first be mapped to a FAO classification for e.g. species, and later, the same set can be mapped to e.g. the species classification of another Regional Fisheries body.
6. In the envisioned VRE, the rdbs of the system would store reference data and mapping rules in tables separate from the actual data tables. Mapping rules will be established using harmonized reference datasets. These data-sets can initially be derived from FAO reference data tables, as described in the ICIS Requirements, but the system should also provide for the storage and management of other mapping schemas.
7. Mapping rules may be exploited across VREs, but only where data-sources and data-sets are released by their manager.
8. In case 1:1 mappings are not possible, more complex mappings can be defined.
9. For each dimension, separate mapping rules can exist, but it is assumed that the dimensions are independent. E.g. area is totally independent from Unit.
10. Once mapping rules are established, performing time-series measures reallocation and transformations become possible. Reallocation is a process which takes as input a statistical dataset (timeseries and related classifications), and produces, by means of established classification mappings and reallocation rules, a new dataset where the measures have been potentially redistributed (depending on the type of mapping) according to another classification.
11. A mapping system also needs to alert when 'matching' code(s) is(are) not available within the coding systems, and needs to allow for new codes to be inserted, with a defined relationship to the acknowledged codes, i.e. either "a part of", "embracing", or "partial overlap".

12. Mapping rules for non 1:1 matching will have to be dealt with either by simple algorithm or manually, e.g. by assigning a ratio. Initially, FAO will develop those rules and inform D4S.
13. A protocol is required on how to manage different versions of datasets, i.e. how to merge newly imported records with existing data. Should existing records be replaced, updated or otherwise be marked. Traceability to sources used is however important, and this requires further discussion.
14. The D4S infrastructure will enable re-allocation of data, according to the mapping rules, and create new datasets from source data. These datasets can hold reference to, but no longer are authorized data (provided by the sources). Users should always be made clearly aware when re-allocations includes data modifications.
15. The infrastructure shall also enable search for data discovery purpose (e.g. which datasets contain statistics about sardine in the Atlantic ocean)
16. The infrastructure shall also support queries for data extraction and comparisons among different time-series, with flexible query terms. Scenarios include: compare across data-sources (e.g. 2000-2005 tunas catch in the North-Atlantic by species between EU data and ICCAT data), compare existing time-series with new data, compare re-allocated results across data-sources (e.g. 2000-2005 tunas catch in Area = {39.97712, 29.443359, 5} by species between EU data and ICCAT data)
17. The infrastructure shall deliver in the first prototype search and data extraction that allow users to search over dimensions, in data ranges (e.g. using <=> operators), and enabling aggregate functions.
18. The reference data should be available in a query builder to search over higher taxonomic levels (e.g. continent search, species groups)
19. The prototype export functionalities should be developed in line with the FishStatJ application and possibly catering for standards (SDMX).
20. Future exchange formats could include e.g. webdav file protocols or data services over http. E.g. a scenario would be to develop a FishStatJ extension that could consume query results from ICIS. This would result in a tight integration of various FAO information sub-systems.