Mapping Data Views in fusion research – a practical approach

G. Manduchi, A. Luchetta, A. Rigoni, C. Taliercio

Consorzio RFX, Corso Stati Uniti, 4, 35127 Padova, Italy

Fusion experiments produce a large amount of data because parameters of interest such as plasma temperature and density are not measured directly, but are derived from possibly large sets of data channels. A medium-size fusion experiment typically involves thousands of data acquisition channels with sampling rates up so several MHz, thus producing very large data sets, especially in recent long lasting discharge devices. Flat databases would be unfeasible for the proper data management and data systems used in fusion research typically provide a hierarchical data organization, representing the first step towards a data ontology, that is, towards a schematic model of data within the specific application. The same considerations apply not only for experimental results, but also in plasma modelling and simulation.

Depending on the application, different perspectives in data are required. For example, a diagnostician requires a data view that reflects the physical organization of his/her diagnostic, but data from that diagnostic, and possibly others, may participate in the data sets involved in a large simulation program. Clearly in this case two different data views are involved and, in this context, tools for defining and exporting different data ontologies are desirable.

Two main aspects must be considered when implementing specific data views:

- Mapping data semantics, that is, defining the data organization that best fits the current use case;
- Metadata integration, that is, enlarging data sets with accessory information required for the proper interpretation of data.

From a practical point of view, the mapping process consists in a definition of a set of classes whose structure best fits the specific application, i.e. is intuitive and efficient. An example is the Interface Data Structures (IDS) defined in the Integrated Modelling Analysis Suite (IMAS) of ITER.

An experimental tool will be presented for the data-driven definition of structures and mappings in order to define data views for the content of experiment-specific pulse databases. The tool has been developed using the Django python framework and uses concepts borrowed by the Semantic Web. Using this tool, the definition of a specific data view is derived in three steps:

- Definition of the data structures (classes) that best fit the specific use case;
- Definition of the actual data items (aka pulse file) using the classes defined in step 1;
- Mapping the atomic data items to actual data sources (experiment databases, metadata definition, simulation analysis).

A Python Application Programming Interface (API) is then generated on the fly, based on the above specifications in order to access data via the specified view.