Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2016) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

13th International Conference on Current Research Information Systems, CRIS2016, 8-11 June 2016, Scotland, UK

R&D Statistics Information System: An Interoperability Tail between CERIF and SDMX

Dimitrios C. Karaiskos[a]*, Dimitrios Xinidis[a], Vasilis Bonis[a]

[a]*National Hellenic Research Foundation/National Documentation Center, 48 Vassileos Constantinou Ave., Athens, 11635, Greece*

Abstract

Research and Development statistics (R&D statistics) provide valuable information on the expenditure spent and personnel engaged in R&D activities in a country, knowledge that facilitates the understanding on how R&D output contributes to economic growth and societal wellbeing. This endeavor requires a sound evidence base which is succeeded through internationally comparable statistics and a common survey methodology and conduct per country as part of its national official statistical program. For this purpose the National Documentation Centre of Greece (NDC or EKT using the Greek abbreviation), the designated organization for the collection and compilation of the Greek R&D statistics, build the R&D Information System to automate and support this specific business activity. This paper aims to provide an overview of the implemented R&D Information System but especially to focus on the adoption of CERIF and SDMX standards and their integration. CERIF was selected as the systems' data model for its metadata representation capability and its high flexibility in forming semantic relationships while SDMX was adopted as the statistical data and metadata exchange standard. The integration of the two standards and their interoperability enables data and metadata quality maintenance, archiving and access while at the same time ensures valid and automated interchange of statistical information with national and international statistical offices.

Peer-review under responsibility of the Organizing Committee of CRIS2016.

Keywords: CERIF; SDMX; R&D Statistics; Systems Interoperability

* Corresponding author. Tel.: +30 210 7273 945
E-mail address: karaiskos@ekt.gr

Peer-review under responsibility of the Organizing Committee of CRIS2016.

1. Introduction

Research and Experimental Development statistics (R&D statistics) provide valuable information on the expenditure spent and personnel engaged in R&D activities in a country, information that is vital for understanding how R&D outputs contribute to economic growth and societal wellbeing. This endeavor requires a sound evidence base which is achieved through internationally comparable statistics and a common survey methodology and conduct per country as part of its national statistical program (based on Commission Regulations 995/2012). In this context, the collection and compilation of the Greek R&D statistics, has been assigned to the National Documentation Centre of Greece (www.ekt.gr) and is supported and carried out by a data management system, namely R&D Information System, implemented in-house specifically for this purpose. The R&D Information System was based on relevant international standards -such as CERIF and SDMX-, robust technologies and best practices. This paper aims to provide an overview of the developed R&D Information System but especially to focus on the adoption of CERIF and SDMX and the integration implemented between them. The next sections will attempt to shed light on the usage of both standards and will describe with detail the integration and interoperability between the two standards.

2. Background

2.1. R&D Statistics

The aim of the R&D (Research and Experimental Development) survey is to produce statistics about (intramural) R&D expenditure and R&D personnel covering R&D performing entities in the private and public sectors and for the country as a whole[1]. The R&D data collection is promoted by EC regulations 995/2012 (from reference year 2012 onwards) and it is carried out on a mandatory basis by each country, each year, according to EU regulations[2, 3, 4]. Subsequently, the country level R&D statistics are relayed to major statistical bodies, such as Eurostat and OECD, who aggregate and compose statistics reports regarding EU and/or the world in various levels of detail. For the R&D statistics aggregation to occur, internationally comparable statistics and a common methodology and language is required. This is provided by the Frascati Manual[1] which outlines all the basic statistical concepts and definitions, standard classifications and guidelines for the production of R&D statistics.

R&D statistics have gained momentum and have increased international interest due to the fact that the financial resources devoted to the implementation and support of R&D activities have been linked to the economic development of a country. A concrete example is the EU2020 headline target of investing 3% of the EU GDP on R&D which is monitored by the 'R&D intensity' indicator (R&D expenditure as a percentage of GDP)[5]. Furthermore, the decision to treat expenditure on R&D as a capital investment in the System of National Accounts (SNA)[6], i.e. accounted positively in national GDPs, has also attracted great attention to the flows of funds for R&D and its robust tracking and measurement. Thus, R&D statistics form the basis for the development and monitoring of policies at European and national levels.

2.2. SDMX Standard

The main purpose of SDMX (an ISO standard ISO 17369:2013 since 2013) is to tackle the problem of a common model for the representation of statistical data and metadata and to provide a set of guidelines and standards for exchanging statistics information between peers, e.g. international organizations and states. The guidelines of SDMX are concerned not only with data but also with metadata which are introduced as an integral part of the standard. Regarding information exchange, SDMX describes the structure of the content and the supported system architectures. It provides the flexibility for each organization that chooses to use it, to structure its data according to their specific informational needs and usage scenarios. The SDMX message formats have two basic expressions, SDMX-ML (using XML syntax) and SDMX-EDI (using EDIFACT syntax and based on the GESMES/TS statistical message). EKT uses the SDMX-ML format since it is the format that Eurostat's web services support and because of the extensive usage of XML in the systems that EKT develops and uses.

The increased popularity of SDMX can be attributed to several reasons. The organizations that sponsor SDMX are undeniably established worldwide and their stature and authority warrants that SDMX will remain a key player in statistics data modeling and exchange. The flexibility and expressiveness of the data model ensures firstly that all informational needs can be met and secondly that there is no need to cut corners in terms of data completeness and quality. Implementing appropriate web services and data management systems becomes a more manageable issue because the available guidelines already describe relative system architectures, thus eliminating the need for an ad hoc design, and also because tools for transforming and exchange data are readily available[7]. For the same reason there is no need for extended discussions for finding a common ground in an organizations consortium in terms of data model and system architecture. Also being an ISO standard ensures the availability of high quality specifications and detail versioning. In the context of EU members, Eurostat will introduce SDMX as the main standard for statistics data exchange so to this end EKT's choice for using SDMX was not only the most appropriate but also obligatory[8].

2.3. SDMX Reference Infrastructure (SDMX-RI)

Eurostat supports EU members to adopt SDMX standard by providing the SDMX Reference Infrastructure (SDMX-RI)[9]. SDMX-RI is a set of open source software tools that enables an organization to quickly adopt SDMX and be able to exchange their statistical data using SDMX formats. By using SDMX-RI the data providers acquire the tools to translate their statistical data into SDMX and to expose it to the external world based on web services' architecture standards (REST API, SOAP). Furthermore, SDMX-RI is designed to provide data and structural metadata based on mappings to the data provider's dissemination database. A dissemination database is the data warehouse (or database) where the data provider (e.g. EKT) maintains statistical data ready for publication to potential data consumers (e.g. Eurostat).

3. Design/Architecture

3.1. R&D Statistics Information System Architecture

The R&D Statistics Information System serves the objectives of a) R&D micro-data collection, b) workflow-based statistical analysis, c) R&D indicators production, d) benchmarking analysis with 3rd party datasets and e) provision of access to R&D statistics to all stakeholders. The desired functionality is achieved by four subsystems, namely the RDI Organisation Registry, the Online Data Collection System, the Data Management System (DMS) and the SDMX Reference Implementation. The RDI Organisation Registry stores information about the sample, i.e. organisations and contact persons, which is surveyed for collecting the R&D micro-data. Also, it is used for managing access control and permissions for all services, including access to the online R&D surveys. The Online Data Collection System is where the R&D questionnaires reside and where the organisations are invited to enter in order to participate. As a modern online survey tool, it covers a multitude of important requirements in conducting a survey, such as real time data validation, participants management and answers management.

The Data Management System, is the single management point for all datasets involved (micro-data, paradata, organization data, and indicators data), thus it is used in order to gather, store, interconnect and manage all the collected R&D micro-data, the organisations' profiles and the produced R&D indicators. Also, it enriches the datasets semantically using appropriate classifications for each dataset. Furthermore, it serves the goals of data preserving and archiving (time series), data validation & estimation workflows, real time automated generation of R&D indicators, data exporting (CSV, Excel, JSON etc.) and statistical reporting. The exporting and reporting capabilities of DMS are used primarily for internal purposes, such as data/survey quality auditing, data benchmarking with 3rd party datasets, publishing of statistical reports, among other. The dissemination of R&D indicators is accomplished through the SDMX Reference Implementation (SDMX-RI) which is the subsystem responsible for making the produced R&D indicators and their metadata available to all interested stakeholders. Stakeholders with accredited permissions to access this information are able through a REST API to query for the R&D indicators and receive them in SDMX-ML format.

3.2. DMS: The core of R&D Statistics Information System

DMS serves a threefold purpose, to provide a software tool for the management of the micro-data that are collected through EKT's R&D annual questionnaire based research, to preserve the data and to help statisticians to produce a set of predefined indicators. In this context, CERIF was chosen as the main data model for the system as it provides the needed data entities, the semantics mechanism, the entities relations and the temporal dimension that are of paramount importance for the success of a system with the characteristics and specifications of DMS. However, minor extensions were deemed necessary for storing the online surveys structure, coding, labels and metadata. The extensions were implemented by adding new entities and appropriately linking them with the relevant existing CERIF entities, while no other changes were made to the CERIF model. The core data that the system manages consist of the various organizations' and institutions' answers that are collected through the R&D questionnaire system. The management of micro-data includes the necessary checks and processing in order to find possible discrepancies and to prepare and produce the indicators that are sent to Eurostat and are included in the publication of relevant studies. The micro-data, the processed data and the indicators are all interlinked and stored under the hood of DMS.

As mentioned before, the micro-data are subjected to a variety of checks before the final statistics indicators are produced. The checks are structured into specific phases (initial state, validation state and estimation state) where the statistician can produce sets of indicators for each phase. The actual validation of data is done in separate third party statistics software packages (such as R and SPSS). The users can initiate separate analysis workflows starting by choosing a subset of initial data. The content of each subset is dictated by various factors like who is the user (different users work with different subsets) and what is the purpose of the analysis e.g. producing indicators for Greek universities only. The results for each user's workflow and for each phase are stored so as to maintain a detailed history of the analysis. One of the final outputs of DMS is a set of indicators and metadata structured in SDMX format as is described extensively in the following sections.

4. Interoperability between CERIF and SDMX

One of the main objectives of R&D Statistics Information System is to provide access to R&D statistics to all interested stakeholders and especially Eurostat. As SDMX is the selected exchange standard format for this purpose, R&D Statistics Information System, and specifically the DMS system, was integrated with SDMX. The integration occurred at both the data and the systemic layer. In particular, at the data layer the SDMX semantic model (the so called DSD) was mapped to CERIF and stored in DMS. Additionally, the R&D indicator's specification (as given by Eurostat) was also mapped to CERIF, stored in DMS and interlinked with the semantic layer of SDMX. The integration at the data layer was the first step to make indicators' values properly available to be exposed by the SDMX-RI. Thus, at the systemic layer DMS was integrated with SDMX-RI, acquiring that way a REST API that is able to receive SDMX queries and to respond with SDMX-ML data messages. The following sections discuss in detail these integration steps.

4.1. Semantic Interoperability: Integrating SDMX DSD to CERIF

The DSD acronym stands for the term Data Structure Definition. It defines the vocabulary used for the characterization and interpretation of the statistical indicators in the context of the SDMX standard. Being a major part of the SDMX protocol, it acts as its semantic layer, providing to the protocol the means to describe and characterize the indicators' values. It contains all necessary description entities, which are bound to the actual indicator's value in order to compose the final result, thus ensuring data robustness and comprehensiveness. The DSD files are distributed in XML format by Eurostat and the purpose of their usage is to describe how indicators' information is structured.

The body of the DSD XML file consists of three main sections, the Concept Scheme, the Code Lists and the Key Family. Their relationships at the entity level are depicted in Figure 1. The Concept Scheme contains the definition of the Concepts, while the Code Lists are predefined sets of terms from which the Concepts retrieve their values. Every Code List contains a number of available values where each one is expressed by the Code entity. The Key

Family section provides the structure between the Concepts and their correlated Code Lists. It also defines the metadata of each Concept as follows:

- *Dimensions* identify the observation value and determine the dataset's physical structure.
- *Attributes* add additional metadata to the observation value.
- *Measure* is the observation value, which is considered a special type and contains the actual value of the indicator.

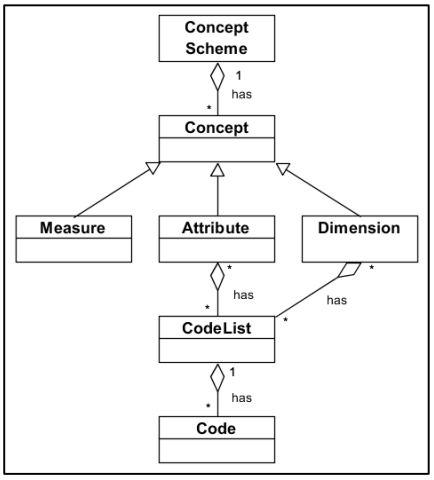

Figure 1: SDMX DSD entities structure

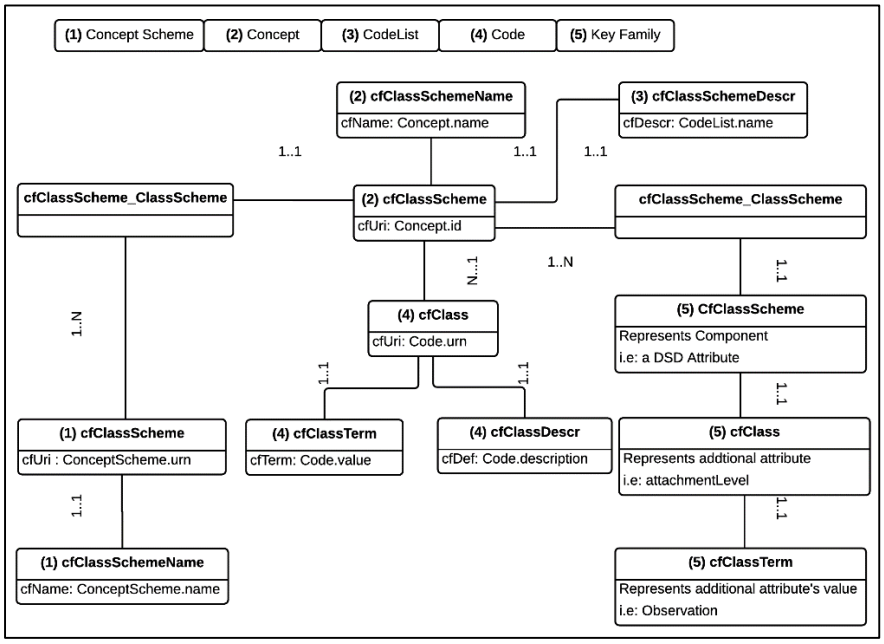

Figure 2: Mapping SDMX DSD to CERIF

As it becomes apparent, DSD is the semantic layer of SDMX standard equivalently to the CERIF semantic layer. This similarity along with the strong expressiveness of the CERIF relational model, facilitates the integration of DSD's data in the CERIF database schema. The migration of DSD data to CERIF is performed by a dedicated mechanism developed in Java, using a generic workflow-based migration tool, called Biblio-Transformation-Engine (BTE)[10]. The migration mechanism depends on the mapping between the DSD and the semantic layer of the CERIF model, as depicted in Figure 2. Specifically, the mapping between the DSD and the CERIF entities is as follows:

- Concept Schemes are mapped as Class Schemes (first level) utilizing the entities cfClassScheme, cfClassSchemeName and cfClassSchemeDescr.
- Concepts are mapped as Class Schemes (second level) utilizing the entities cfClassScheme and cfClassSchemeName. Concepts are linked to their relevant Concept Schemes using the ClassScheme to ClassScheme link entity.
- Code Lists are stored at Class Scheme Description (second level) as they provide the grouping metadata of the Codes that compose each Concept.
- Codes are mapped as Classifications utilizing the entities cfClass, cfClassTerm and cfClassDescr.
- Key Family provides the structure between the Concept Schemes, Concepts, Code Lists and Codes and it is realized through the appropriate linking entities. However, it was decided mainly for maintenance reasons to be also stored explicitly in the database using the cfClassScheme, cfClass and cfClassTerm entities.

4.2. Semantic Interoperability: Integrating R&D Indicators to CERIF

Similarly with the DSD, Eurostat provides the specification of the R&D indicators as an accompanying file to the DSD. The R&D indicators are grouped in several categories and carry notations, labels and semantics that are important for their calculation and presentation via the DMS. In fact, the specification file contains for each indicator a) its calculation formula, b) its defining categories and labels and c) its metadata, i.e. which concepts and codes from the DSD are relevant with the specific indicator. The contents of the R&D indicator's specification file are processed and stored into DMS by a dedicated procedure implemented for this purpose. The procedure is divided into four main parts as follows:

- The indicator part, responsible for defining the indicator identity (name, category, description etc.)
- The subindicator part, which is the breakdown of each indicator based on the calculations according to the indicators' specifications file. At the subindicator level is where the actual measurement values are calculated and stored in DMS. Each indicator includes multiple subindicators, i.e. the composition of a numerous related subindicators synthesizes the final indicator outcome.
- The mapping of each subindicator and its related metadata to the DSD entities, i.e. Concept Scheme, Concept and Code.
- The connection of the indicators' data flow with the existing DMS analysis' data flow. The indicators' values are calculated over the raw data that for statistical purposes are grouped under a single entity called analysis.

Figure 3 shows how the indicators, subindicators, measurements and the metadata accompanying them are mapped to CERIF entities and stored in DMS. Once the procedure is completed, the indicator flow has been constructed and connected to the rest of DMS data flow and to the related dictionary scheme (DSD). At that time the system is considered ready to manipulate the indicators' values and produce reports to the end users by integrating each indicator value with its related descriptive metadata.

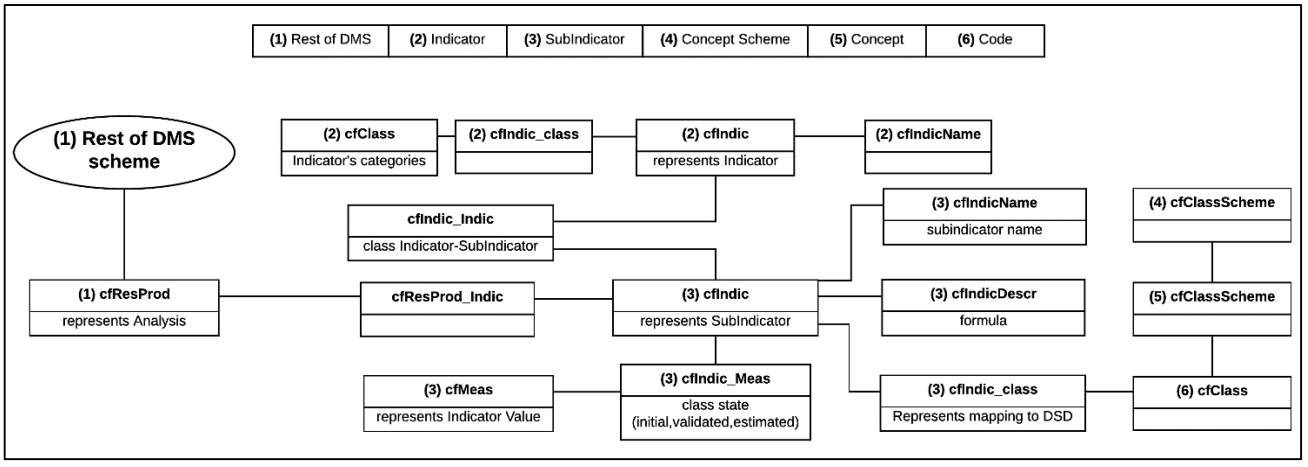

Figure 3: Mapping between CERIF and R&D indicators

4.3. System Interoperability: Integrating SDMX-RI to DMS

SMDX-RI through its web service interface provides an SDMX gateway for a dissemination database to the outside world. This database should conform to several patterns in order to be compliant to the SDMX transformation scheme. First it should contain the DSD vocabulary definition in order to form the SDMX structures. Additionally it must hold the real indicators' data, described as SDMX data. The web service provider will create the final SDMX-ML message, as a combination of the SDMX's structure and data. This combination is implemented based on a configuration mapping structure, created by the SDMX Mapping Assistant[11], an application

provided by Eurostat. Its task is to fill the Concepts defined by the DSD file with the required data. This is accomplished by connecting each DSD Concept to the metadata or the indicator's data depending on the Concept's type as defined in the DSD file under the Key Family section. DMS has been designed to support multiple DSD files for different scopes and also for adjustments and revisions of existing DSD files throughout the years. Using the same philosophy the SDMX Mapping Assistant maintains a new mapping structure for every new DSD definition file imported in DMS in its internal database schema.

Since the DMS database schema is based on CERIF, which is a very fine grained normalized and complex model, the structural and indicator data are extracted in a form of a graph data structure with multiple nodes and vertices. The SDMX Mapping Assistant provides the option to use custom and sophisticated queries on top of the extracted data and thus allows an indirect map for more complex datasets such as graphs. Though, due to the large number of indicator's data and the CERIF's model complexity, such queries require multiple levels of abstraction to become suitable for mapping, resulting in a major decrease on the overall system's performance. This is the reason why the option of using a classic two-dimensional table as the data feed for the SDMX Mapping Assistant was preferred. The idea behind this is that the indicators' data and metadata will be provided to SDMX Mapping Assistant in a simplified form (two dimensional). To implement this, an extra layer of data processing was injected on top of the DMS schema acting as a façade for the SDMX infrastructure. This transformation logic applies a post processing procedure in order to transform the data in an SDMX compliant format, by placing the performance burden on the side of the DMS. This burden is transparently absorbed by DMS itself, providing an adequate overall performance for the system. Figure 4 depicts the entire flow triggered by an SDMX compliant client.

The Web Service provider receives the client's request and it channels it through the following modules:

- The *Sdmx Parser* parses the incoming request message.
- The *Data Retriever* interacts with the DMS dissemination and mapping store databases to extract the indicators data in a suitable format.
- The *Structure Retriever* interacts with the DMS dissemination and mapping store databases to extract the metadata in a suitable format.
- The *SDMX Data Generator* packs the responses of the Retriever modules in an SDMX-ML format and sends them back to the client, which initiated the request.

The same figure shows the "transformation" layer, which sits on top of the DMS's data layer and transforms the CERIF data to an SDMX compliant format, ready for processing by the Data Retriever and the Structure Retriever modules. The mapping store, as described above, contains all the mappings that are performed from the context of the SDMX Mapping Assistant.

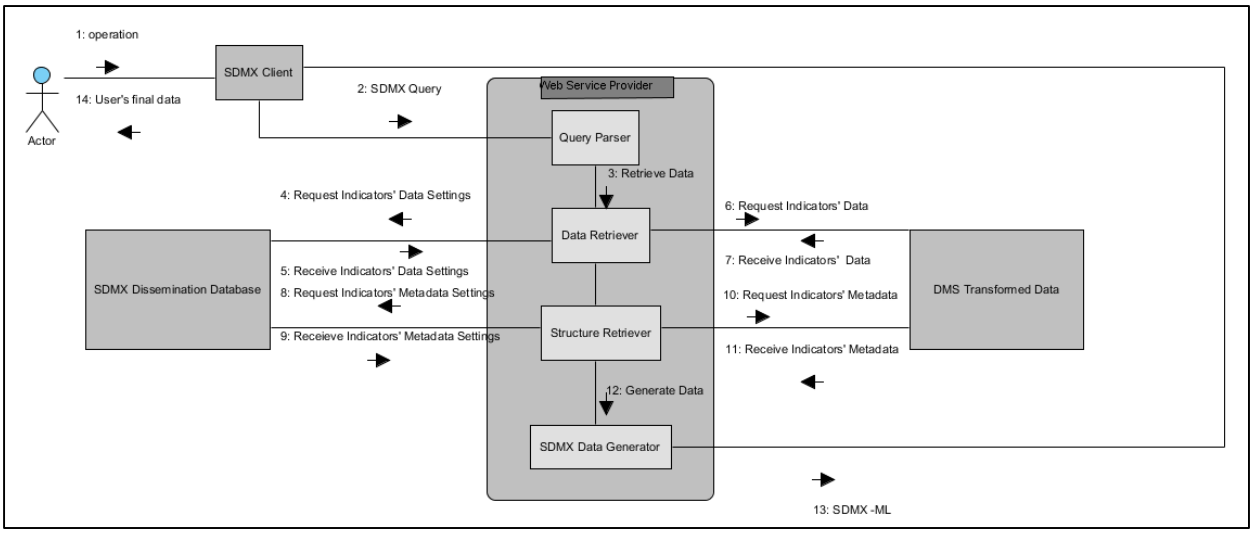

Figure 4: Activity flow between SDMX-RI and DMS

5. Conclusions and further steps

The discussion about R&D statistics is certainly not something new. Internationally the activities of collecting, archiving, processing and exchanging relative information have proven their value as a vital tool for understanding the research and academic environments. Furthermore, it is evident that research is and always will be a fundamental framework on top of which a country's macroeconomic figures can thrive. EKT through its formal role has invested on creating the appropriate IT infrastructure for facilitating the aforementioned activities by taking into account not only national parameters and needs but also international best practices, methodologies and standards. CERIF and CRIS systems have long been fields of vivid engagement for EKT and they were quickly recognized as useful and flexible tools for supporting this type of IT systems. Having a single point of focus for these activities has been proven very important as it simplifies the data inputs and workflows, the processing components and the information outlets.

A CERIF based system, as demonstrated by DMS, is capable of handling the related information and facilitating the needed processing mainly due to CERIF's flexible and robust semantics structure. Although DMS is not strictly speaking a CRIS system, one can argue that since CERIF can be used for the aforementioned purposes, a CRIS can also play a significant role in creating an information network for R&D statistics. The collection and archiving of source data of relative themes have already been distinct features of CRIS systems. The DMS paradigm described in this paper can be used as a basis for implementing the IT processing tools and SDMX through CERIF can handle the exchange of data. This information network can effectively demonstrate a natural flow of information from the sources to the consumers.

The decision to adopt CERIF and SDMX provides the appropriate tools to empower the shift to a national R&D data infrastructure of greater granularity down to the level of R&D projects, involved organizations, researchers and performed activities. The standards system put in place through the Frascati Manual has helped to trace R&D funds across sectors at an aggregate level but this in turn has stimulated demand for more disaggregated information[12]. Thus, it is recognized that a higher degree of granularity is required to track funding, knowledge, activities and people in order to define how R&D funding drives change. Towards this end, our future steps involve the integration of DMS with the National R&D Projects database aiming to gradually gain the ability to benchmark the collected R&D aggregated data over relative funding as objectively depicted through R&D funded projects.

References

1. OECD, Frascati Manual: Proposed Standard Practice For Surveys On Research And Experimental Development, 2002.
2. E. Commission, "Commission Implementing Regulation (EU) no 995/2012," Official Journal of the European Union, 26 October 2012.
3. E. Commission, "Decision no 1608/2003/EC of the European Parliament and of the Council," Official Journal of the European Union, 22 July 2003.
4. E. Commission, "Commission Regulation (EC) no 753/2004," Official Journal of the European Union, 22 April 2004.
5. E. Commission, EUROPE 2020: A strategy for smart, sustainable and inclusive growth, 2010.
6. U. N. S. Commission, "System of National Accounts," 2008. [Online]. Available: unstats.un.org/unsd/nationalaccount/docs/SNA2008.pdf.
7. "SDMX Tool Repository," [Online]. Available: www.sdmxtools.org/index.php.
8. Eurostat, "SDMX Newsletters and implementation projects," [Online]. Available: webgate.ec.europa.eu/fpfis/mwikis/sdmx/index.php/SDMX_Newsletters_and_implementation_projects.
9. Eurostat, "SDMX reference infrastructure (SDMX-RI)," [Online]. Available: ec.europa.eu/eurostat/web/sdmx-web-services/sdmx-reference-infrastructure-sdmx-ri-.
10. EKT, "EKT/Biblio-Transformation-Engine," [Online]. Available: github.com/EKT/Biblio-Transformation-Engine.
11. Eurostat, "Mapping Assistant," [Online]. Available: webgate.ec.europa.eu/fpfis/mwikis/sdmx/index.php/Mapping_Assistant.
12. OECD, Working Party of National Experts on Science and Technology Indicators, Proposal For Oecd/Nesti Work On The Proof Of Concept For An Analytical International Micro Database On Public R&D Project Funding (Fundstat), Paris, 17-18 March 2016.