**UNECE High-level Group for the  
Modernisation of Official Statistics**

**Project Proposal: CSPA Data Architecture**

|  |
| --- |
| This project proposal was prepared by Modernization Committee on Standards / Juan Muñoz, INEGI, and is submitted to the HLG-MOS for their approval. |

|  |  |
| --- | --- |
| **1 Purpose** | |
| To develop a reference framework for data architectures for the official statistics industry, as a part of the Enterprise Architecture which is one of the fundaments of the design of the Common Statistical Production Architecture (CSPA). | |
| **2 Project description** | |
| The project consists of the development of a reference framework, to describe a standardized data platform to support the design, integration, production and dissemination of official statistics.  The definition given by TOGAF (The Open Group Architecture Framework – an international standard for enterprise architecture) for a data architecture could be a guide for defining the main deliverables of the project: “*A description of the structure and interaction of the enterprise's major types and sources of data, logical data assets, physical data assets, and data management resources”.*  The project outputs will be:   * A description of the structure and interaction of the major types and sources of data. This deliverable can be interpreted as a reference model describing a concrete set of the core data structures needed to support the statistical information production, in relation to the Generic Statistical Business Process Model (GSBPM) * Guidelines to describe conceptual artefacts like statistical data dictionaries and data catalogues to drive the definition of data structures and metadata * A standardized catalogue of the common logical clusters of data which are relevant to statistical organizations * Guidelines and recommended practices for managing these described data assets to ensure sufficient data quality for statistical organizations to collaborate and share technical solutions (like the CSPA services), and knowledge, such as: comparability, interpretability, transparency, interoperability, fluidness, integrity, consistency, and so on. * Architectural Information Capabilities Guide describing capabilities that statistical organizations need to efficiently and effectively implement future-proof data and metadata architectures   The architecture will consider traditional and emerging data sources (such as Big Data, geo-spatial data, etc.) along with metadata. This means that it will not only cover just static data structures but dynamic data objects too. The CSPA Data Architecture will describe data sources, like:   * Structured databases * Structured and unstructured files with different formats * Streams of data coming from emerging data sources like social networks, sensors, etc. * Information objects described by the CSPA Logical Information Model (CSPA LIM) and exchanged by CSPA statistical services * References to remote data sources   This project is considered as a fundamental step to enable efficient data and metadata management and governance in the context of CSPA. It supports and builds ideas from the Modernisation Committee on Production and Methods about “Next Generation Data Management”. It has been defined by the need to satisfy new and more sophisticated demands of information products and services, where this only can be achieved making use of all kinds of data sources, traditional and emerging. | |
| **3 Alternatives considered** | |
| 1. Extending the CSPA information architecture section to include some specifications and principles. However, these are too related to the needs of service architects and developers from the point of view of an application layer. A data architecture goes beyond supplying information to the applications, and must provide the definitions for an institutional data platform that avoids data “silos” (which prevent the flow of information between areas, and limits the ability to use and to integrate this data to create new products that satisfy user needs). 2. Enhancing the scope of the CSPA LIM, to provide more instruments to implement a data architecture (and not only the definitions of the information flows between statistical services). This approach is not recommended because it will change the purpose defined for the CSPA LIM, creating confusion and making it more complex. Nevertheless, CSPA LIM as a logical layer could be very useful for providing a guide to help to define the data architectures. 3. Doing nothing. This is not an option, because it limits the potential for sharing CSPA services, by not having similar information contexts and integrated data platforms. An Enterprise Architecture is based in four layers (business, data, applications and technology), all of which are needed. | |
| **4 Expected Benefits** | |
|  | Reduced costs |
|  | Increased efficiency |
|  | Reduced risks |
|  | New capabilities to meet user needs |
| Justification:  Some expected benefits from the project are:   * Enhancing governance of data inside a statistical organization * Eliminating the existence of internal data silos, reducing redundancy and inconsistency of information and making information flows between statistical domains more efficient by reducing internal and external obstacles * Helping to guide the data strategy of the statistical organization, by aligning the efforts of its different units * Driving the definition of data structures and metadata, for collection, reception, integration, processing, dissemination and exchange of information * Making it easier to source data from different domains, for new products and services to satisfy new users’ needs * Addressing information management challenges preventing statistical organizations from quickly leveraging new data sources * Defining how to incorporate and integrate traditional and emerging data sources (such as geo-spatial data and Big Data) * Making it easier to integrate and to manage information from different kind of sources, not only traditional databases but also unstructured and dynamic data * Reducing time to market to produce new statistical products and services * Making it easier to use existing standards, like DDI and SDMX, among others * Helping to better understand of the data and metadata that are managed by statistical organizations * Enhancing the ability of statistical organizations to share information and knowledge * Enhancing the comparability and transparency of the information held in different statistical organizations * Helping to adopt common CSPA statistical services, by providing a common data context in which they can operate, and paving the way to having an integrated statistical information production platform * Positioning statistical organizations to leverage opportunities based on technological developments in linked data and metadata, next generation data storage, analysis, visualisation, new methods and algorithms | |
| **5 Which key priorities in the HLG-MOS Strategic Framework does the proposed project relate to?** | |
|  | Take cost out of our organisations to reinvest in more value added areas |
|  | Explore new areas collectively and leverage each other’s' research investments in specific areas |
|  | Provide whole of government data ecosystems based on international standards, for better estimates in key policy areas |
|  | Renew our governance and operating processes |
| Justification:   * It reduces re-work needed to deliver information to different applications * It helps in the adoption of different standards, by providing a common data context where they can be implemented * It supports the development and implementation of the data governance needed to support the alignment of efforts around a data strategy * It reduces costs, by helping to share information, applications and knowledge with less effort, and using common tools | |
| **6 How does the proposed project relate to other activities under the HLG-MOS?** | |
| It complements and consolidates the CSPA approach, by taking care of the most important assets for statistical organizations, data and metadata.  It applies the business framework provided by GSBPM, and makes concrete use of GSIM and CSPA LIM standards.  It covers the layer between business process and applications defined by an enterprise architecture. | |
| **7 Proposed timetable** | |
| The project will run from January to December 2017.  Project progress and draft outputs will be reviewed by the Executive Board as appropriate.  Project outcomes and outputs will be presented at the Workshop on the Modernization of Official Statistics in November 2017. | |
| **8 Expected resources and costs** | |
| 20 person months spread across an estimated 10 participating organizations and the UNECE Secretariat  6 person months for a project manager  Costs associated with 10-12 persons attending a physical sprint session in early 2017 and a project meeting back-to-back with a suitable international event such as the 2017 CSPA Implementation Workshop. | |