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

**Business case for the Core Ontology for Official Statistics Project**

|  |
| --- |
| This business case was prepared by Insee and is submitted to the HLG-MOS for their approval. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **1 Purpose** | | | | |
| As more and more statistical offices turn to semantic standards in order to formalize their data and metadata, it is time to build on our international core models to establish common foundations on which the different works can develop in a coherent way. This should be done using a formal framework that allows model integration, interoperability, activation and globally unique identification.  Initial work on a base OWL vocabulary for Official Statistics has begun. It suggests formal representations for the core concepts and models used in the official statistics domain, and builds on a set of well-known OWL vocabularies, which favors interoperability with similar initiatives in other domains (finance, legal, industry, etc.).  This project aims to consolidate this material in order to create a common semantic model and vocabulary for official statistics. If successful, the project's output will form the base for the next generation of HLG standards. | | | | |
| **2 Description of the activity** | | | | |
| Various works have already been conducted in order to represent the GSBPM, the GSIM, parts of CSPA and the SDMX metadata model in OWL (see <http://ceur-ws.org/Vol-1551/article-06.pdf>, <http://ceur-ws.org/Vol-1654/article-03.pdf>, <http://ceur-ws.org/Vol-1654/article-06.pdf> and <https://linked-statistics.github.io/SDMX-Metadata/sdmx-metadata.html> respectively). They show that OWL provides an excellent tool for the integration of the models into a coherent and interoperable framework.  Additional work was pursued at Insee in order to create a base ontology covering the core notions of statistical or support activities, as well as statistical organizations and products. In terms of existing standards, the ontology covers the GSBPM and the GAMSO, most of the GSIM base package, and the CSPA organizational roles. The objectives of this work were:  - to create a common ground to which the more sectorial works listed above could be attached to form a coherent set of vocabularies  - to provide at this higher level the relevant connections with well-known external standards (namely SKOS, PROV, ORG, DCAT and its application profiles like StatDCAT-AP, the Dublin Core, the Data Quality Vocabulary, etc.), in order to benefit from their rich semantics and from all the developments based on them (for example dataset catalogues conformant to DCAT).  Building on this initial work, further work is needed:  Activity 1: Provide precise definitions of the core ontology concepts and their properties, and define their identification in relation to a clearly documented naming policy. This activity also includes the specification of the links between this base ontology and the external vocabularies mentioned above and others that could also seem appropriate.  Activity 2: Articulate this base ontology with more sectorial and in-depth works existing on GSIM, CSPA or the SIMS model for quality metadata (based on the SDMX metadata model). This could lead to marginal evolutions of these models. The identification of the resources defined in these works will also be adapted in order to align with the naming policy defined in Activity 1.  Activity 3: Establish the management and governance of these different initiatives. The management activity should be tightly integrated with the management of the current HLG-MOS standards, but also deal with the connection to external vocabularies and corresponding standardization bodies.  The project will produce the following deliverables:   * A URI policy for UNECE RDF resources * The Core Ontology as an OWL/Turtle file in English/French/Spanish, validated by public review and referenced in [LOV](http://lov.okfn.org/dataset/lov/). * An ontology description and users guide (English) * New versions of previous works on GSIM, CSPA and the SDMX metadata model * A document on the governance   A tentative schedule for the project is provided below:  *January*  Constitution of a group of reviewers and liaison with "Supporting Standards" group  *February*  Draft naming policy  *March*  Internal review  *April*  Draft version 1 of the ontology after internal review  Start of public review  *May-June*  Work on articulation with work on GSIM and CSPA  *July-August*  Work on governance  *September*  Final version 1 of the ontology  *October*  Presentation in the SemStats workshop  Document on governance  *November*  Submission to HLG-MOS for adoption | | | | |
| **3 Alternatives considered** | | | | |
| The following alternatives were considered:  1. Do nothing  In this alternative, different NSIs will invest in duplicate work which will produce incoherent results. No globally recognized semantic modeling or naming policy for the base semantics of official statistics will be available for the community to build on.  2. Realize the project outside of UNECE  In this alternative, the work is carried out under a different umbrella (standards organization like the W3C or the DDI Alliance, international interoperability frameworks like the ISA2 European program, etc.). In this case, there is a great risk that the results will not be clearly articulated with the HLG standard models, and in particular that it will be hard to coordinate their evolutions with those of the UNECE standards. There is also a clear risk that the results will not be fully accepted by the UNECE or its members. | | | | |
| **4 Expected Benefits** | | | | |
| ☐ | | | Reduced costs | |
| ☒ | | | Increased efficiency | |
| ☐ | | | Reduced risks | |
| ☒ | | | New capabilities to meet user needs | |
| Justification:  - the possibility of representing various metadata (e.g. discoverability, provenance, quality, etc.) according to well-known standards like PROV or DCAT, which is provided by the links between these standards and the base ontology, will improve the visibility of our projects and the ease of access to them  - the anchoring on these standards will also foster interoperability with semantic models developed in other sectors, for example the FIBO ontology in the financial domain  - the existence of a common base ontology will improve the interoperability of the developments within the statistical community  - the activity results will provide guidance for the statistical organizations and prevent multiple reinventions of the wheel | | | | |
| **5 Type of Activity** | | | | |
| ☒ | | New activity | | |
| ☐ | | Extension of existing activity | | |
| ☐ | | Other *(specify below)* | | |
|  | | | | |
| **6 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 | | | |
|  | | | | |
| **7 How does the proposed activity relate to other activities under the HLG-MOS?** | | | | |
| The proposed activity is tightly linked to the HLG-MOS activities related to standards, in particular the GSBPM, the GAMSO, the GSIM and CSPA.  The proposed activity also builds on work completed in the 2016 HLG project: *Implementing Modernstats Standards: Linked Open Metadata*.  It is also related to the Data Architecture project, which prescribes the use of OWL for the semantic layer of the architecture, and of linked metadata in particular in the "metadata management" and "provenance and lineage" capabilities. | | | | |
| **8 Proposed start and end dates** | | | | |
| **Start:**  **End:** | | | | January 2018  December 2018 |
| **9 Expected costs** | | | | |
| It is estimated that the work will need the involvement of a group of experts on statistical models and knowledge representation, under the coordination of a project manager. Project management can be estimated as a half-time job for the duration of the project.  Travel costs for a three- to five-day sprint should also be anticipated. | | | | |