HLG Project: Implementing ModernStats Standards

Towards a unified implementation and global integration of HLG-MOS and other global models and standards, and providing a Roadmap for their implementation.

Rationale

Modernisation is a key priority for all statistical organisations. The HLG-MOS has set out a vision defining long-term goals and has overseen the development of the necessary models and standards to help statistical organisations modernise. Developing a Modernisation Roadmap and providing modernisation standards as linked open metadata are logical next steps towards implementing the HLG-MOS vision.

This project will:

- Respond to growing demands to provide key modernisation and metadata standards and models[1] in machine-readable formats[2] to support data integration and growing pressure from governments and other users in many countries to provide open data. This will encourage greater use of these standards within official statistics, as well as in the wider data community. This will facilitate data sharing, as well as the use of new and emerging data sources for statistical purposes.
- Provide a central repository of key standards in the form of "linked open metadata" reducing the need for statistical organisations to develop such a resource individually. This will increase consistency in implementations, and will enhance understanding of linkages and dependencies between standards and models, facilitating more efficient maintenance. Several statistical organisations, for example CBS, INSEE and Istat, have already started to invest in linked open metadata and linked open data, hence this project comes at the right time to share experiences and emerging best practices
- Develop a Modernisation Maturity Model, for statistical organisations to determine their current level of maturity in terms of readiness to apply modernisation standards. This will be complemented by a Modernisation Roadmap, providing clear guidelines on the steps to take to reach higher levels of organisational maturity. This roadmap will be based on the experiences and lessons learned from the more advanced statistical organisations, so will help others to progress more quickly and efficiently.

Introduction

The statistical community develops metadata standards of good quality, but these standards are rarely available for internal or external users in open and machine-actionable formats. The linked data format is especially relevant for metadata dissemination, because:

[1] E.g. GSBPM, GSIM, GAMSO, Classifications such as ISIC, ISCED etc., key concepts and definitions
[2] The UNECE Secretariat has already received a request from Statistics Finland for GSIM in a machine-readable format.

- global uniform naming and addressing is crucial for structural metadata such as classifications, code lists, cube dimensions, etc.;
- there exist several linked data standard models or vocabularies which are dedicated to metadata[3],and allow for good discoverability and referencing by external users;
- glossaries and vocabularies expressed as linked data can be connected to or from other concept schemes such as Eurovoc or the Library of Congress Subject Headings (LCSH);
- It will enlarge the base of users of statistical data through better integration with standards from outside the statistical community.

There is a clear demand from the academic world for reference linked statistical metadata (concepts, code lists, etc.): this was strongly expressed at the recent Semantic Statistics (SemStats) workshop, for example.

Beyond dissemination, formalizing metadata with linked data standards guarantees a level of coherence, interoperability and adaptability that other models do not offer. For example, semantic descriptions of CSPA services can be formally linked to the GSBPM sub-process in which they operate and to GSIM objects that form their inputs and outputs. It is also in line with the active metadata paradigm, since linked data are easy to consume automatically and to integrate into metadata-driven processes. Thus, linked metadata are also a promising tool for achieving a better consistency and integration, within each organization and at the global level, of the statistical production.

A growing number of statistical institutes have understood that problem and have started to invest in linked metadata, so now is a good time to share experiences on the subject at an international level. The benefits that can be achieved include:

- Easier data access, in terms of obtaining and combining data in machine-readable formats;
- Quality tracking by comparing, reproducing, finding inconsistencies and facilitating correct interpretation;
- Integration by linking data, avoiding unnecessary duplication, ensuring semantic homogeneity (same concept/variable, same name), searching multiple database at once for combining data.

In addition, the European Statistical System is launching the DIGICOM project, one of the key activities of which is linked data. 2016 will be a year of stocktaking for this activity, so it is important to acquire experience on the subject.

Objectives

The project has two main objectives. The first is to demonstrate the usefulness of linked metadata for the statistical community and to acquire hands-on experience in that field. It is proposed to fulfil this objective by constructing two concrete examples of linked metadata-based information systems: one aimed at improving the way that we disseminate core

[3]For example the Dublin Core, SKOS/XKOS, PROV or DCAT.

structural metadata, the other at supporting the advancement of the HLG-MOS vision by creating a harmonized and semantically enhanced information system grouping the main models and standards in a coherent and machine-actionable form. Each of those systems constitutes a deliverable of the project.

The second main objective is to provide the means for statistical organisations to evaluate their levels of maturity against a standard framework with the aid of a Modernisation Maturity Model, and to help them determine the priorities for next steps based on a roadmap.

Scope

Regarding the first main objective, the project aims at demonstrating the interest of linked metadata for the Official Statistics community by developing example systems corresponding to two different use cases.

The first system illustrates a dissemination use case: the goal is to conceive and build a database of globally harmonized core statistical metadata (concepts and codes) and to make it available in both user-friendly and machine-actionable formats. The tools and methods used to create the system will be documented and made available so that statistical organizations can produce additional content and link it to the core system. The query, extraction and visualisation tools will also be shared.

The second use case has a more internal focus: the aim is to create a semantically enhanced information system supporting the HLG-MOS vision. This system will contain semantic representations of the main business models published by the HLG-MOS and of the CSPA standards for documenting services. The idea is to show the benefits of a unified representation of the models and services, as well as to develop functionalities going from a better ease of access to the possibility of linking additional contents (translations, process descriptions, etc.) to the system.

In addition to the development of the two demonstrators described above, the project will produce a sustainability plan so that the HLG-MOS can make an informed decision as to whether to ensure the continued existence of the systems after the end of the project.

The second main objective aims to address the requirements expressed by a number of statistical organisations, particularly those in the earlier stages of modernization, to have clearer information about how to progress along the road towards modernising statistical production in the most efficient way. The MMM and Modernisation Roadmap should however also be useful to any organisation regardless of their level and acknowledge that within an office they can be at different level depending on the domain or activity.

Content

The project will have four work packages. The first two will be focussing on showcasing the practical usage of linked statistical metadata. Each activity has clearly defined output driven deliverables. The second work package will do so by unifying the main HLG-MOS models. The third work package builds on this by developing a roadmap for implementing these

HLG-MOS models and standards. Finally, the fourth work package will deal with the overall project management and evaluation, and will define a sustainability plan for the maintenance of the output of the other work packages.

Work Package 1: Build a dissemination system for core structural metadata

This system will contain a core set of international structural metadata (concepts, classifications and codes) and propose querying, extraction and visualisation functionalities both for human and machine consumption. The tool will implement the CSPA "Classification" service description.

Activity 1.1 Specify the content and its representation

This activity will include the selection of the metadata sets to represent as linked data. The following are expected to be included:

- Glossaries (UNECE metadata glossary and, if possible, SMDX glossary)
- International classifications (at least ISIC, NACE, CPC, CPA)
- International standard code lists (to be selected)

If possible, additional international classifications, as well as correspondences between the core classifications, will be also incorporated.

This will also cover the specification of the RDF vocabularies and URIs that will be used to represent and identify the metadata.

Deliverable 1.1 - Content specifications: metadata list, vocabularies, URI schemes

Activity 1.2 Create the linked metadata set

Deliverable 1.2.1 - Documented programs to create the RDF files

Deliverable 1.2.2 - RDF files containing the selected metadata sets

Activity 1.3 Create a demo tool to query and visualize the metadata

Deliverable 1.3.1 - Functional and technical requirements

Deliverable 1.3.2 - RDF database with query interface

Deliverable 1.3.3 - Web visualisation interface

If possible, additional tools will be developed, for example for converting the query results in different office or database formats.

Activity 1.4 Showcase of usage of the system in statistical offices dissemination

Deliverable 1.4.1 - Description of practical usage of the system in current dissemination activities of NSIs

Work Package 2: Build an information system supporting the HLG-MOS vision

The implementation of the standards in a coherent framework, where the inter-relationships among standards are made more evident as opposed to implementing each standard alone, will enhance the synergies among the standards. The tool to be developed will contain a

semantic coherent representation of the main HLG-MOS models (GSBPM, GAMSO, GSIM, CSPA service descriptions). It will allow to browse the information (e.g. clickable GSBPM or GAMSO) and to specify new information based on a number of use cases to be chosen in a first phase.

Possible use cases could be: providing the translation of a model's text content into another language, defining national refinements of a model, attaching specific process descriptions to GSBPM sub-processes, assisted documentation of CSPA services, linking external models like the ESS SPRA, etc.

Activity 2.1 Specify the system

Deliverable 2.1.1 - Selection of use cases (model dissemination, translations, refinements, service descriptions, linking, etc.)

Deliverable 2.1.2 - Data model: representation of the main HLG-MOS models (GSPM, GAMSO, GSIM, CSPA service descriptions) as linked data

Activity 2.2 Build the system

Deliverable 2.2.1 - RDF file containing the metadata sets selected

Deliverable 2.2.2 - RDF database with query interface

Deliverable 2.2.3 - Web interface implementing the use cases selected

Work Package 3: Modernisation Roadmap

This work package will provide the means for statistical organisations to evaluate their levels of maturity against a standard framework with the aid of a Modernisation Maturity Model, and to help them determine the priorities for next steps based on a roadmap.

It aims to address requirements expressed by a number of statistical organisations, particularly those in the earlier stages of modernization, to have clearer information about how to progress in the most efficient way. These requirements have been expressed in several forums, including the Workshop on Modernisation of Official Statistics, Nizhny Novgorod, June 2014 and the Workshop on Standards-based Modernisation in Geneva, May 2015.

The Modernisation Maturity Model (MMM)

The MMM will be an evaluation tool to assess the level of organizational maturity against a set of pre-defined criteria. The development of the MMM will be led by the Modernisation Committee on Organisational Frameworks and Evaluation, but will be open to other interested parties.

A good starting point seems to be The Open Group Service Integration Maturity Model (OSIMM), which defines a maturity model in the following way: "A maturity model is a means of and scale for evaluating and assessing the current state of maturity. A maturity model also provides a means for developing a transformation roadmap to achieve a target state of maturity from a given current state of maturity. It quantifies the relative growth of certain salient aspects within various dimensions typically within, but not limited to, organizational boundaries."

The Modernisation Committee on Production and Methods successfully used a modernisation maturity model developed by the Australian Bureau of Statistics, based on OSIMM, in a pilot study to assess readiness for adopting software components (services) developed in compliance with the

Common Statistical Production Architecture (CSPA). This pilot will be studied to determine the extent to which this approach could be applied more widely.

Maturity is indicated by the attainment of a particular "maturity level". A maturity level assessment will provide a clear picture of the organisational maturity level, which can be compared between organisations, and between processes within an organisation. It will be based on a consistent set of questionnaires and an independently held set of "benchmarks".

The Modernisation Roadmap

Knowing the level of organizational maturity is an obvious pre-requisite to designing activities to increase capabilities in relevant areas. However, it is also necessary to understand the optimal sequence of capability development activities, based on the experiences and lessons learned of the organisations that have already progressed to higher levels of organizational maturity. In this way, statistical organisations that are less advanced can progress more quickly, without repeating the mistakes others may have made.

The roadmap will focus on building organizational capabilities through implementation of the models and standards identified as key to statistical modernisation, such as The Generic Statistical business Process Model (GSBPM), and its extension the Generic Activity Model for Statistical Organisations (GAMSO), the Generic Statistical Information Model (GSIM), the Common Statistical Production Architecture (CSPA), including the CSPA Logical Information Model currently under development. The roadmap will also indicate types of support that statistical organisations, at different maturity levels, need in order to implement the different standards.

Key external standards such as SDMX (Statistical Data and Metadata eXchange) and DDI (Data Documentation Initiative) are also important for statistical modernisation, and are clearly connected to both the GSIM and the CSPA. These, and other relevant standards, will also be mentioned as examples of best practice, to help to increase the level of maturity, but they are not the primary focus of this project.

Activity 3.1 Agree on the scope and a set of dimensions along which modernisation maturity will be measured and create a draft MMM

Deliverable 3.1: draft MMM

Activity 3.2 Test and update the MMM

Deliverable 3.2.1: A trial of the draft MMM amongst the participants in the project

Deliverable 3.2.2: Updated version that can be used by all statistical organisations

Activity 3.3 Create a roadmap to guide organisations on how to implement the standards (GSBPM / GSIM / GAMSO / CSPA) in the context of the MMM

Deliverable 3.3: A roadmap, indicating paths and milestones on the road to implementation of the standards (GSBPM / GSIM / GAMSO / CSPA) in the context of the MMM

Interest in participation for WP.3

Already 11 statistical organisations have indicated their interest in participating in this work package: Australian Bureau of Statistics, Statistics Estonia, Hellenic Statistical Authority in Greece**,** Central Statistics Office in Ireland, Central Bureau of Statistics in Israel, Istat in Italy, Statistics Finland, INSEE in France**,** National Institute of Statistics and Geography of Mexico (INEGI), Statistics Norway, Office for National Statistics in UK.

Work Package 4: Project management, evaluation and sustainability plan

The evaluation will describe the results achieved with regard to the project's objectives, the possible problems encountered and the lessons learned that can be useful for the NSIs or for other similar initiatives (for example DIGICOM). This evaluation will provide the matter for a presentation of the project's results that will be given at the HLG-MOS 2016 workshop but can also be made for different stakeholders and on other occasions.

The tools described in the previous sections will be available for demo at HLG-MOS 2016 workshop. It is necessary to propose a plan (organization, governance, updating, funding) to make the linked metadata and the HLG-MOS standards support system durable. It is equally important to provide a platform to provide guidance and to facilitate collaboration for implanting the modernisation roadmap.

Activity 4.1 Project management and evaluation

Deliverable 4.1.1 Overall Project planning and periodic project evaluation

Deliverable 4.1.2 Provide guidance and induction of project participants

Deliverable 4.1.3 Ensure timely delivery of deliverables WP.1-WP3.

Deliverable 4.1.4 Project presentation

Activity 4.2 Define a sustainable mechanism for the maintenance of the project outputs

Deliverable 4.2.1 Sustainability plan for maintenance of HLG-MOS linked statistical metadata and its availability

Deliverable 4.2.2 Proposal for MMM and Roadmap support portal to guide statistical organizations in implementing key standards for modernisation

The plan will make a distinction between the two systems corresponding to WP.1 and WP.2, since the measures for sustainability can be different in each case, and the conclusions of the HLG-MOS can also be different. Support for MMM and Roadmap includes adjusting them to incorporate modifications and new standards and further includes suggestions for ensuring collaboration and assistance between NSOs to reach the next step in their modernisation.

Definition of success

In the most restrictive perspective, the project will be a success if the different deliverables described under WP.1 and WP.2 are completed on time and if the MMM under WP.3 is developed and promoted by the international statistical community.

In a broader view, the project will meet success if the HLG-MOS endorses the sustainability plan and decides to ensure the durable existence of one or both of the information systems created in the framework of the project. Real success would further mean that statistical organisations are able to use the roadmap to help them to move to a higher maturity level and that they are supported in implementing the standards in a coherent and comprehensive framework/approach.

An even bigger achievement would be that the project fulfils its overall objective of convincing the official statistics community that linked metadata are an effective paradigm for the modelling, management and dissemination of statistical metadata, and that they can provide powerful leverage for the modernisation of the statistical process. Furthermore, that the MMM would be referred to and used as one of the best practices required for the modernisation process to achieve efficient production of high quality official statistics, oriented to satisfy the needs of its users.

Expected costs and risk factors

Open and free tools will be used for development and demo platform, so most of the costs consist of human resources. Some travel costs have to be expected, especially for development sprints.

The evaluation of the costs associated to long term availability of the systems will be provided in the WP.4 deliverables (D4.2.1 and 4.2.2).

Work Package	Resources	Source	Other costs
WP.1	12 person months per activity	Volunteer NSOs (10pm), UNECE Secretariat (2pm)	Possible travel costs Workshop/Sprint
WP.2	12 person months per activity	Volunteer NSOs (10pm), UNECE Secretariat (2pm)	Possible travel costs Workshop/Sprint
WP.3	12 person months	Volunteer NSOs (10pm), UNECE Secretariat (2pm)	
WP.4 Project Management	6 person months	A part-time project manager from a volunteer NSOs working with the UNECE Secretariat Input from Executive Board and HLG-MOS members (in their role as project sponsors)	Travel costs for project events and telecommunications and other incidentals
Total	42 person months		US$

Risks that can be identified for the project include under-estimation of the conceptual or technical difficulties, obstacles to the availability of the data to be included in the systems or lack of coordination with relevant actors: other modernization committees (on Standards, on Production & Methods), other organizations or projects (UNSD, Eurostat and DIGICOM). Developing the Roadmap will also depend on the prior or joint development of the MMM.

By aiming at having members of these groups actively involved in the project, and by communicating regularly about the project and its progress, risks will be minimized or mitigated where possible.

Time lines

The project will aim to complete the activities described by the end of 2016. The three work packages will start simultaneously and run for the full period of the project. For work packages 1 and 2, the specifications deliverables should be completed at the end of the first trimester and the linked databases should be available by the end of May. The specific activities in WP.3 will be sequential according to the following scheme:

Activity	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
3.1	X	X	X									
3.2				o	o	X						
3.3		o	o	o	o	o	o	o	X	X	X	

o - low activity, X - high activity

Governance

The project sponsor is the HLG-MOS. This is the group that has ultimate responsibility for signing off the project deliverables. In practice, this responsibility will be delegated to the Executive Board. They will be supported by the four HLG-MOS Modernisation Committees.

The Modernisation Committee on Products and Sources will advise on WP.1 and WP.2. In particular, it will make the decisions relative to the exact content of the systems (deliverables 1.1 and 2.1.1). The Modernisation Committee on Standards will provide input on the standards to be used in the first two packages and specifically on how the main HLG-MOS models inter-relate (Activity 2.1). Likewise, the Modernisation Committee on Production and Methods will guide WP.3 by providing its experiences and the Modernisation Committee on Organisational Frameworks and Evaluation will lead the development of the MMM.

The project manager under WP.4 will have day-to-day responsibility for the running of the project, providing regular updates and signalling any issues to the Executive Board as necessary.