HLG project proposal on Linked Statistical Metadata

# Introduction

HLG has been jointly developing common models and vocabularies to prevent each organization from developing their own using different vocabularies for the same concepts[[1]](#footnote-1). Linked open metadata provides the next step. Instead of each organization having to maintain and update their individual vocabularies, this would be made available and managed in a centralized way. This not only reduces costs but also prevents discrepancies in structural and reference metadata and semantic heterogeneity.

These are only the start of the benefits. Linked open metadata are also the key enablers of linked open data as they allow for accessing and sharing data across organizations. In open data settings, it allows for searching and integrating data from a large number of data sources and combined with the semantic web, to endless possibilities of associating it with (contextual) information available from outside structured databases.

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

* global uniform naming and addressing is crucial for structural metadata like classifications, code lists, cube dimensions, etc. so that there is a unique and well-known reference always accessible by standard mechanisms;
* there exist several linked data standard models or vocabularies which are dedicated to metadata[[2]](#footnote-2), which allows for good discoverability and referencing by external consumers or publishers;
* glossaries and vocabularies expressed as linked data can be connected to or from other concept schemes available on the internet in this format, like [Eurovoc](http://eurovoc.europa.eu/) or the Library of Congress Subject Headings ([LCSH](http://id.loc.gov/authorities/subjects.html));the linked data format enlarges the base of users of statistical data by relying on standards also coming 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 problematic 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 are several and various, including:

* Easiness of data access, in terms of fetching the data and combining data in whatever format/concept scheme and machine readability;
* Quality tracking by comparing, reproducing, finding inconsistencies, correct interpretation,
* Integration by linking data, avoiding unnecessary duplication of data, 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 main objective of the project 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 structural metadata, the other at supporting the advancement of the HLG vision by creating an harmonized and semantically enhanced information system grouping the main CSPA models and standards in a coherent and machine-actionable form.

Each of those systems constitutes a deliverable (and thus a secondary objective in itself) of the project. Other important deliverables will be a project evaluation describing the lessons learned and documenting best practices, and a sustainability plan for the projects outcomes beyond the end of 2016.

## Scope

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 High Level Group vision. This system will contain semantic representations of the main business models published by the HLG 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 can make an informed decision as to whether to ensure the continued existence of the systems after the end of the project.

## Content

The project will have three work packages. Each activity has clearly defined output driven deliverables.

WP1 will first define a set of metadata (glossaries, classifications and codelists) as candidates for the linked metadata dissemination system. Such metadata will be implemented as linked metadata sets, with a specific attention on documenting the implementation process and on explaining design choices. A prototype tool will demonstrate how to query and visualize the linked metadata sets. In addition, practical usages of the linked metadata set in NSIs dissemination activities will be described in detail.

WP2 has the goal of unifying the main HLG models (GSPM, GAMSO, GSIM, CSPA service descriptions) in a coherent representation. To the purpose, first a linked data representation of such models will be provided, then a system will be built containing such representations, with a dedicated work on documenting design and implementation choices. Finally, a prototype tool will show the usage of such representations with respect to selected use cases.

WP 1 and WP 2 could possibly share the same technological platform for both linked data generation and prototyping activities on accessing and visualizing generated linked data.

#### 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.

Activity 1.1 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 vision

This tool will contain a semantic coherent representation of the main HLG models (GSPM, GAMSO, GSIM, CSPA service descriptions). It will allow to browse the information (eg. 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](http://www.cros-portal.eu/content/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 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: Project 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 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 2016 workshop. It is necessary to propose a plan (organization, governance, funding) should the HLG decide to make the system durable.

Activity 3.1 Project evaluation

Deliverable 3.1.1 Project evaluation

Deliverable 3.1.2 Project presentation

Activity 3.2 Define a sustainable organization for the availability and maintenance of the systems

Deliverable 3.2.1 Sustainability plan

The plan will make a distinction between the two systems corresponding to work packages 1 and 2, since the measures for sustainability can be different in each case, and the conclusions of the HLG can also be different.

## Definition of success

In the most restrictive perspective, the project will be a success if the different deliverables described previously are completed on time.

In a broader view, the project will meet success if the HLG endorses the sustainability plan and decides to ensure the durable existence of one or both if the information systems created in the framework of the project.

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.

## Expected costs and risk factors

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

The cost for WP3 only includes the cost of establishing the sustainability plan. The evaluation of the costs associated to long term availability of the systems will be provided in the WP3 deliverables (D3.2.1).

| **Work Package** | **Resources** | **Source** | **Other costs** |
| --- | --- | --- | --- |
| WP1 | 10 person months per activity | Volunteer NSOs plus UNECE Secretariat | Possible travel costs Workshop/Sprint |
| WP2 | 10 person months per activity | Volunteer NSOs plus UNECE Secretariat | Possible travel costs Workshop/Sprint |
| WP3 | 3 person months | Volunteer NSOs plus UNECE Secretariat |  |
| Project Management | 4 person months | A project manager working with the UNECE Secretariat.  Input from Executive Board and HLG members (in their role as project sponsors) | Travel costs for project events |
| Total | 27 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).

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

Several statistical institutes have already started to invest in linked metadata, hence there can be a systematic sharing of experience to face with technical difficulties.

## 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.

## Governance

The project sponsor is the HLG. 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.

The Modernisation Committee on Products and Sources will act as the steering committee for the project. In particular, it will make the decisions relative to the exact content of the systems (deliverables 1.1 and 2.1.1).

A project manager 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.

1. Such as GSPM, GAMSO, GSIM, CSPA service descriptions, classifications of Statistical Activities and of Types of Big Data [↑](#footnote-ref-1)
2. For example the Dublin Core, SKOS/XKOS, PROV or DCAT. [↑](#footnote-ref-2)