Modernstats at UNECE

ESS Modernisation Workshop,
Bucharest 16 – 17 March 2016

UNECE Statistical Division
Taeke Gjaltema
taeke.gjaltema@unece.org
Content

- UNECE
- HLG-MOS/ModernStats
- Main Outputs/Standards
- Projects
- Modernisation Committees
- Other
UNECE & CES

- UN Economic Commission for Europe
  - 56 member countries
  - Europe, Caucasus & Central Asia, North-America

- Conference of European Statisticians:
  - 65+ members
  - UNECE & China, Japan, Mongolia, South-Korea, Australia, New Zealand, Brazil, Chile, Colombia, Mexico, South-Africa, United Arab Emirates

- High Level Group on the Modernization of Official Statistics (HLG-MOS)
Introducing the HLG-MOS

- High-level Group for the Modernisation of Official Statistics
- Created by the Conference of European Statisticians in 2010
- Vision and strategy endorsed by CES in 2011/2012
Rapid changes in the environment

New competitors & changing expectations

Increasing cost & difficulty of acquiring data

Reducing budget

SDGs

Riding the big data wave

Competition for skilled resources
We have to produce more and more statistics, more frequently and more disaggregated with more and more competition but with less resources

➡️ We have to make the statistical production faster, more efficient and use new data sources (admin, big data, geospatial etc.)

➡️ Only a modern, efficient and flexible organization can do this

➡️ Collaboration needed
Statistical Modernization Community

- A vision for an aligned and collaboratively led community
- Allows all of us to benefit from collaboration and sharing
- Four main principles:
  - Openness
  - Flexibility
  - Participation
  - Pragmatism
Structure HLG-MOS

Conference of European Statisticians

HLG-MOS

Executive Board

Moderisation Committees
- Organisational Framework & Evaluation
- Production & Methods
- Products & Sources
- Standards

CSPA Implementation Group

Projects
HLG, What and Who?

- Oversees activities that support modernisation of statistical organisations
- Stimulates development of global standards and international collaboration activities
- “Within the official statistics community ... take a leadership and coordination role”

Members:

- Ireland (Chair)
- Australia
- Canada
- Italy
- Netherlands
- New Zealand
- Rep. of Korea
- Slovenia
- Hungary
- Eurostat
- OECD-SD
- UNECE
HLG-MOS Mission and Roadmap

- Stimulate development of global standards and oversee international collaboration activities
  - Take a leadership and coordination role
  - Collaboration of the willing
  - Focused on delivery tangible/practical outputs
- Implementing common standards and models for the official statistics “industry”
- Promoting collaboration and sharing
  - From the design stage, not just the outputs
- Modular systems giving increased flexibility for new sources / processes / outputs
HLG-MOS Activities – Engagement Map
HLG Achievements:

- **Generic Statistical Business Process Model** (GSBPM)
- **Generic Statistical Information Model** (GSIM)
- **Common Statistical Production Architecture** (CSPA)
- **Generic Activity Model for Statistical Organizations** (GAMSO)
- Big Data: Sandbox, quality, partnerships
GSBPM: GSIM describes the information objects and flows within the statistical business process.

Input
- Any GSIM Information Object(s) (e.g. Data Set, Variable)
- Process parameters

GSBPM
Sub-process

Output
- Transformed (or new) GSIM Information Object(s)
- Process metrics
Mappings

Fundamental Principles of Official Statistics
**GAMSO**

- **Generic Activity Model for Statistical Organisations**
- It extends and complements the GSBPM by adding other activities needed to support statistical production
CSPA

Fostering Interoperability in Official Statistics: Common Statistical Production Architecture
Big Data HLG-Project:

- **2014 Project:**
  - Guidelines on Partnerships, Privacy, Quality and Skills
  - IT / methodological issues - **Sandbox**

- **2015 Project:**
  - Sandbox
  - **Big Data Inventory**

- **2016 Sandbox:** continued collaboration
  - Subscription model launched

- The Irish Centre for High-end Computing hosts the UNECE Big Data ‘sandbox’ containing data and tools for international experiments

“Play is the highest form of research”
– Einstein

UNECE Big Data Wiki:
http://www1.unece.org/stat/platform/display/bigdata
Big Data Sandbox: 2016

- More sandbox experiments
- Continue collaboration
- Concrete outputs
- Future of the sandbox
  - Subscription model launched
  - Not just for Big Data – other uses e.g. shared development of methods, shared training materials, …
Projects and Modernisation Committees

- HLG Projects 2014-15:
  - Big Data
  - CSPA

- HLG Projects 2016:
  - Data Integration
  - Linked Statistical Metadata

- Modernisation Committees:
  - Standards
  - Production and Methods
  - Products and Sources
  - Organisation Framework and Evaluation
  - CSPA governance and maintenance
MC-Organisation Framework and Evaluation

- Change management and risk management
- Legal and licensing (HLG statement of intent)
- Building competencies: survey, big data skills
- Guidelines for managers
- Modernisation Maturity Model & Roadmap
MC-Production and Methods

- Support to CSPA Projects
- Workshop on the Modernisation of Statistical Production / Data Editing
- Machine learning
- Generic Statistical Data Editing Models
- Methodology architecture
MC-Products and Sources

- Big Data Inventory
- Communication: surveys value of official statistics
- Workshop on Statistical Communication and Dissemination & Workshop on Statistical Data Collection
- Mobile devices
- Data integration/mixed mode/linked statistical metadata/administrative data
MC-Standards

- Metadata glossary
- Roadmap for implementing standards (HLG project)
- Quality indicators for GSBPM
- Further development of GSIM concepts / variables
- Logical Information Model
- GAMSO
- Workshop on International Collaboration for Standards-based Modernisation
The story so far - 2015
What’s now and next in HLG-MOS

- GAMSO usage
- Logical Information Model
- HLG projects:
  - I) Data Integration
  - IIa) Linked statistical metadata
  - IIb) Modernisation Maturity Model & Roadmap
- Modernisation Committees:
  - Many activities
HLG Implementing ModernStats Standards

- Linked Statistical Metadata, important:
  - Have a globally unique identification and location for metadata
  - Model metadata in relation to existing standards
  - Enable linking metadata
  - Ensure machine-actionability in order to activate metadata

- Modernisation Roadmap:
  - Roadmap for implementing standards (GSBPM / GSIM / GAMSO / CSPA) in the context of a modernisation maturity model
HLG Linked Statistical Metadata project

- Make statistical standards available
- Unified representation of HLG standards
- Key enablers of linked open data
- Searching and integrating data from multiple sources
- Integrate with Semantic web
Modernisation can mean different things to different people depending on their starting point. The level of maturity will vary across organisations but also within organisations across domains.
Modernization Maturity Model

- We need to know how ready we are to adopt new ideas and solutions
- We need to understand the logical sequence of modernisation activities
- We can collaborate with those at a similar level, and learn from those that are ahead of us
  - but how do we know who is at what level?
Implementing ModernStats Standards

- **WP1 – metadata dissemination**
  - Activity 1.1 Specify the content and its representation
  - Activity 1.2 Create the linked metadata set
  - Activity 1.3 Create a demo tool to query and visualize the metadata
  - Activity 1.4 Showcase of usage of the system in statistical offices dissemination

- **WP2 – HLG models**
  - Activity 2.1 Specify the system (use cases, data models
  - Activity 2.2 Build the system (RDF database, implementation of selected use cases)

- **WP3 – Modernisation Roadmap**
  - Activity 3.1 Agree on the scope and a set of dimensions along which modernisation maturity will be measured and create a draft MMM
  - Activity 3.2 Test and update the MMM
  - 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
HLG Data Integration Project

“We must move from a paradigm of producing the best estimates possible from a survey to that of producing the best possible estimates to meet user needs from multiple data sources”

(Conny Citro)

Produce **stable output** with **unstable** ever changing **inputs**

**Opportunities**
- Big data and administrative data
- New technologies

**Challenges**
- Increasing needs: timeliness, frequency and disaggregation
- Less budget, lower response burden
- Other statistics producers

It’s Time to Collaborate!
HLG Data Integration project

- Integrating survey, admin, big data, non official data
- Common data sets to collaborate in sandbox
- Methodology, Metadata & Quality frameworks
HLG Data Integration project

- WP-0: prepare common data sets in sandbox
- WP-1: Integrating Survey and Administrative Sources
- WP-2: New Data Sources and Traditional Sources
- WP-3: Integrating Geospatial and Statistical Information
- WP-4: Micro-Macro Integration
- WP-5: Validating Official Statistics
- WP-A: Methodologies and quality frameworks: Synthesis Lessons Learned from new working methods & draft framework

- “Placeholder” workpackages
- Actual work in concrete sub-projects
  - Depending on interest and feasibility
- HLG standards like GSBPM, GSIM will be used and if necessary, updated
UNECE vs Eurostat

- Membership UNECE wider:
  - CIS countries, Asia, Latin America
- Other governance
- UNECE more bottom up, Eurostat (perceived) as more top down
- UNECE more opt-in opt-out basis
HLG-MOS & ESSNet

- Coordination to prevent duplication
- Synergy to complement projects
- Cross participation NSO staff
- Eurostat staff member/participants HLG, MCs & HLG projects
Welcome to the UNECE Virtual Standards Helpdesk

This wiki provides a "one-stop shop" for access to information about the standards necessary for the modernisation of official statistics. It is an initiative of the High-Level Group for the Modernisation of Statistical Production and Services (HLG).

The standards referenced here are cross-cutting, supporting the modernisation of all types of statistical production, and are endorsed by the HLG. For domain-specific standards, please see the Global Inventory of Statistical Standards.

Generic Statistical Business Process Model (GSBPM)

The GSBPM describes and defines the set of business processes needed to produce official statistics. It provides a standard framework and harmonised terminology to help statistical organisations to modernise their statistical production processes, as well as to share methods and components.

Generic Statistical Information Model (GSIM)

The GSIM is a reference framework of information objects, which enables generic descriptions of the definition, management and use of data and metadata throughout the statistical production process. It provides a set of standardized, consistently described information objects, which are the inputs and outputs in the design and production of statistics. As a reference framework, GSIM helps to explain significant relationships among the entities involved in
Get involved!
Anyone is welcome to contribute!

More Information
- HLG Wiki: www1.unece.org/stat/platform/display/hlgbas
- LinkedIn group: “Business architecture in statistics”
Extra Slides

- GSBPM
- GAMSO
- CSPA
- Big Data
- MC members
- MMM
- Mode of work
What is the GSBPM?

- A flexible model that describes and defines the set of business processes needed to produce official statistics
- Standard framework and harmonised terminology to help statistical organisations
  - Modernise statistical production processes
  - Share methods and components
Why do we need the GSBPM?

- To define and describe statistical processes in a coherent way
- To compare and benchmark processes within and between organisations
- To make better decisions on production systems and organisation of resources
Process-oriented approach

- Statistical production has traditionally been organised by topic, e.g. transport, trade, …
- Financial pressures are encouraging new ways of thinking
- Some statistical organisations are moving towards a process-based approach
- Others are considering a matrix approach
Structure of the Model

- National implementations may need additional levels
- Over-arching processes
  - Quality management
  - Metadata management
  - Statistical framework management
  - Statistical programme management
  - ......... (others – see GSBPM documents)
Key features

- Not a linear model
- Sub-processes are not followed in a strict order
- It is a matrix, through which there are many possible paths
- Not prescriptive but descriptive
5.4 Edit & Impute
- Impute for missing units or variables
- Edit "unusual" data according to pre-defined edit rules

5.5 Derive new variables & units
- Profiling

5.6 Calculate weights

5.7 Calculate aggregates
- Population totals

5.8 Finalise data files
- Maximize register quality prior to creating outputs

The dotted lines show the different routes depending on whether the output is a sampling frame or a statistical product.
Uses of the GSBPM

- Managing statistical programmes
- Cost / resource allocation
- Documenting statistical processes
- Framework for quality assessment
- Sharing statistical software

Used by more than 50 statistical organizations
Uses of GAMSO

- Resource planning
- Measuring costs
- Assessing readiness to implement different aspects of modernisation
- Supporting risk management systems
- Implementing enterprise architecture
- Measuring and communicating the value of statistical modernisation activities
Maintaining GSBPM & GAMSO

- Owner = High-Level Group
- Maintenance is delegated to the Modernisation Committee on Standards
- Discussion forums to gather feedback
- Importance of stability over time
  - Reviews every 5 years
  - Revisions only if really needed
What next for GAMSO?

- Statistical organisations starting to use it
- Collect feedback
- Review in early 2016
- Revision (if needed!)
GAMSO More information

- Virtual Standards Helpdesk

http://www1.unece.org/stat/platform/display/VSH
CSPA provides a template architecture for official statistics, describing:

- **What** the official statistical industry wants to achieve
- **How** the industry can achieve this, i.e. principles that guide how statistics are produced
- **What** the industry will have to do, compliance with the CSPA
Modernisation blueprint exists

CSPA provides a reference architecture to help each agency modernise based on common standards:

- GSBPM
- GSIM
- DDI / SDMX

CSPA allows us to modernise our environment and use existing international solutions.
Historically, statistical organizations have produced specialized business processes and IT systems.

Many statistical organizations are modernising and transforming using Enterprise Architecture.
When countries work on their own...

....Sharing becomes difficult!

CSPA enables sharing
A Statistical Service

GSIM object structures (formats)

GSBPM - process

GSIM object instances
Sharing Infrastructure across Statistical Organisations

By sharing infrastructure development, we can:

- Reduce costs of development
- Adopt new methods quickly
- Increase comparability of statistics
CSPA Services built

1. Seasonal adjustment – France, Australia, New Zealand
2. Confidentiality on the fly – Canada, Australia
3. SVG generator – OECD
4. SDMX transform – OECD
5. Sample selection business registers – Netherlands
6. Linear error localisation – Netherlands
7. Linear rule checking – Netherlands
8. Error correction – Italy
9. Classification retrieval service
10. Probabilistic record linkage
11. Webdissemination service

Freely available to any statistical organisation
Knowledge Base

Investment Catalogue
Built in 2015
Focus on **future** sharing opportunities

Capability Catalogue
Built in 2015
Focus on **existing** sharing opportunities

CSPA Service Catalogue

Technical Repository
How do I find out more?

CSPA Wiki

http://www1.unece.org/stat/platform/display/CSPA
Big Data
Executive Board, Modernisation Committees

Partners

1 Project Manager
2 Coordinators

3 Task Teams

7 Sandbox Experiment Teams

75 Individuals from 25 countries / organisations
## Big Data tools used Sandbox 2015

<table>
<thead>
<tr>
<th>Tool</th>
<th>W</th>
<th>Twitter</th>
<th>RHadoop</th>
<th>Elasticsearch</th>
<th>Spark</th>
<th>Python</th>
<th>R</th>
<th>Graph visualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hive</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Pig</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>RHadoop</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Elasticsearch</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Spark</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Python</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>R</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Graph visualization</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>
Results 2014
2015 Main activities

- **Wikistats** - *Wikipedia hourly page views*: use of an alternative data source

- **Twitter** - Social media data: experiences comparison in tweets collection and analysis

- **Enterprise websites**: the Web as data source - web scraping and business registers

- **Comtrade** - *UN global trade data*: use of Big Data tools on a traditional data source
Twitter

882,007 Twitter Users
43’079,312 Geo-Refenced Tweets
August 2014
Twitter sentiments

Estado de ánimo de los tuiteros en México

Índice = 🌟🌟🌟
Webscraping job vacancies

![Bar chart showing the accuracy of scrapped data compared to survey data by activity of companies.](image)
Comtrade data

Trade of intermediate goods by geographical region
MC-Organisation Framework and Evaluation

- Jackey Mayda (Chair), (Canada)
- Donna Nicholson and Julianne Jex (Australia)
- Marie Creedon (Ireland)
- Milena Grassia, Antonio Ottaiano, Fabrizio Rotundi (Istat) and Prof. Alessandro Hinna (University of Rome, Tor Vergata)
- Igor Mocanu, (Moldova)
- Carina Fransen, Wouter Jan van Muiswinkel (Netherlands)
- Ingvild M. Moller (Norway)
- Anna Borowska (Poland)
- Thana Chrissanthaki (Eurostat)
- Jonathan Challener (OECD)
MC-Production and Methods

- Rune Gløersen (Norway) (Chair)
- Gillian Nicoll / Michael Meagher (Australia)
- Claude Poirier / Rob McLellan (Canada)
- Elaine Lucey (Ireland)
- Alessio Cardacino (Italy)
- Matjaz Jug (Netherlands)
- Rosemary McGrath (New Zealand)
- Janusz Dygaszewicz (Poland)
- Jan Jones (United Kingdom)
- Hubertus Cloodt / Pal Jancsok (Eurostat)
- Bruno Urban / David Barraclough (OECD)
- Valentin Todorov (UNIDO)
MC-Products and Sources

- Barteld Braaksma (Chair) (Netherlands)
- Julie Trépanier (Canada)
- Franck Cotton (France)
- Eoin McCuirc (Ireland)
- Stefano De Francisci / Monica Scannapieco (Italy)
- Jean Watt (New Zealand)
- Anna Dlugosz (Poland)
- Luxolo Lengs / Koketso Moeng (South Africa)
- Martina Hahn (Eurostat)
- Yuri Gyomai (OECD)
MC-Standards

- Klas Blomqvist (Chair) (Sweden)
- Al Hamilton (Australia)
- Alice Born (Canada)
- Guillaume Duffes (France)
- John Dunne (Ireland)
- Marina Signore (Italy)
- Juan Muñoz (Mexico)
- Adam Brown (New Zealand)
- Jenny Linnerud (Norway)
- Nilgün Dorsan (Turkey)
- Dan Gillman (United States)
- Arofan Gregory (DDI Alliance)
- Alessandro Bonara (ECB)
- Jan Planovský (Eurostat)
- David Barraclough (OECD)
Capabilities – the key to the MMM?

- Capabilities require a combination of organisation, people, processes, methodology, standards and technology.
Activity / Process / Capability

- Activity is what we do
- Process is how we do it
- Capability is what allows us to do it
Virtual working

- We use Webex and wikis
- Flexibility
- Free access
  - From office, home, airport etc.
Virtual sprint

Guidelines

Task teams

Consultation

Physical sprint