HLG - Big Data Sandbox for Statistical Production

Learning to produce meaningful statistics from big data

Tatiana Yarmola
(ex) Intern at the UNECE Statistical Division
INEGI, December 3, 2013
Big Data

• Big Data: volume, velocity, variety
• Big Data potential:
  – more relevant and timely statistics
  – can reproduce a variety of statistical indices ➝
    • reduce survey expenses
    • reduce response burden
  – And more

http://vimeo.com/62289901

12-Dec-2013
Tatiana Yarmola, (ex) Intern UNECE Statistical Division.
High Level Group

• High Level Group for the Modernisation of Statistical Production and Services (HLG)
  – Created by the Conference of European Statisticians in 2010
  – 10 heads of national and international statistical organisations

12-Dec-2013

Tatiana Yarmola, (ex) Intern UNECE Statistical Division.
The Challenges

- Increasing cost & difficulty of acquiring data
- New competitors & changing expectations
- Rapid changes in the environment
- Reducing budget
- Riding the big data wave
- Competition for skilled resources

Tatiana Yarmola, (ex) Intern UNECE Statistical Division.
These challenges are too big for statistical organizations to tackle on their own

We need to work together

Who was involved in GSIM?

Get involved!

Anyone is welcome to contribute!

More Information

HLG Wiki:

http://www1.unece.org/stat/platform/display/hlgbas

LinkedIn group “Business architecture in statistics”
Standards-based Modernisation

2013 Project

GSBPM
Reviewed in 2013

Fostering Interoperability in Official Statistics:
Common Statistical Production Architecture

Mapped in 2013

Reviewed in 2013

Tatiana Yarmola, (ex) Intern UNECE
Statistical Division.
Big Data:
The new project for 2014
What does “Big data” mean for official statistics?*

- **Sources:** administrative, commercial, from sensors and tracking devices, social media, mobile telephones etc.

- **Challenges:** legislative, privacy, financial, management, methodological, technological

- NSO’s have infrastructures to address **accuracy, consistency, and interpretability**

- 3 broad areas of **experimentation:**
  - Combining Big Data with official statistics
  - Replacing official statistics by Big Data
  - Filling new data gaps with Big Data

*http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170614
Collaboration and Activities on Big Data

• **The Big Data Project** specifies key priority areas to be tackled as a collaborative activity in agile fashion
  – No systems in place  ➔ develop collaboratively
  – Common issues ➔ common solutions
  – Learn to address current phenomena
  – Agree on common classification of Big Data types
  – Use administrative resources experience

• Mechanism for sharing: **inventory of big data activities**
  [http://www1.unece.org/stat/platform/display/msis/Big+Data+Inventory]
Big Data Project

- **Objectives**: guidance, feasibility, facilitation of sharing
- **Scope**: the role of Big Data in modernisation of official statistics
- **Work package 1**: Issues and methodology
- **Work package 2**: Shared computing environment (‘sandbox’) and practical applications
- **Work package 3**: Training and dissemination
- **Work package 4**: Project management and coordination

http://www1.unece.org/stat/platform/display/msis/Final+project+proposal%3A+The+Role+of+Big+Data+in+the+Modernisation+of+Statistical+Production
WP2: The Big Data Sandbox

- To produce strong, well-justified and internationally-applicable recommendations on appropriate tools, methods and environments for processing and analyzing different types of Big Data
- To report on the feasibility of establishing a shared approach for using Big Data sources that are multi-national or for which similar sources are available in different countries
- To enable efficient sharing of tools, methods, datasets and results
Sandbox Task Team

Goals of this group:

• Define precisely what concept(s) is/are to be proved
• Clarify the value of the exercise in terms of the HLG and international work
• Explore alternative scenarios regarding choice of
  ▪ tools, environment
  ▪ datasets
  ▪ statistics to be produced
  ▪ methods of producing them
• Outline costs, timeline for work during 2014.
• Explore funding options, including infrastructural support and expertise for the project during 2014.

Deliverables: a detailed specification to present as an annex to the project proposal to the HLG at their November meeting.

<table>
<thead>
<tr>
<th>Name</th>
<th>Country/organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tomaž Špeh</td>
<td>Slovenia</td>
</tr>
<tr>
<td>Carlo Vaccari</td>
<td>Italy</td>
</tr>
<tr>
<td>Martin Ralphs</td>
<td>UK</td>
</tr>
<tr>
<td>Tanya Yarmola</td>
<td>UNECE</td>
</tr>
<tr>
<td>Andrew Murray</td>
<td>Ireland</td>
</tr>
<tr>
<td>Pilar Rey del Castillo</td>
<td>Eurostat</td>
</tr>
<tr>
<td>Vittorio Perduca</td>
<td></td>
</tr>
<tr>
<td>Marco Puts</td>
<td>Netherlands</td>
</tr>
<tr>
<td>Brian Studman</td>
<td>Australia</td>
</tr>
</tbody>
</table>

Tatiana Yarmola, (ex) Intern UNECE Statistical Division.
Deliverables

• web-based Big Data Sandbox + specifications + virtual machine releases + training materials

• relevant datasets

• proof of concept:
  datasets $\iff$ manipulate $\iff$ statistic

• common methodology of producing statistics from Big Data

• communicate results

Tatiana Yarmola, (ex) Intern UNECE Statistical Division.
Example of a proof of concept

Figure 4. Dutch consumer confidence (grey) and the overall sentiment in Dutch social media messages on a monthly basis (black). Dutch articles are used as search terms. The social media sentiment in December is considerably more positive compared to the sentiment in the months before and after.

“Big Data and Official Statistics” by Piet Daas et al.
Platform and Tools Criteria

• Ease and efficiency
• Open source or low cost
• Ease of integration with other tools
• Integration into existing statistical production architectures
• Availability of documentation
• Availability of online tutorials and support
• The existence of an active and knowledgeable user community.
Platform Recommendations

• Processing Environment: **HortonWorks Hadoop distribution**

• Processing tools/software: **The Pentaho Business Analytics Suite Enterprise Edition**

  • Pentaho Business Analytics Suite Enterprise Edition provides a unified, interactive, and visual environment for data integration, data analysis, data mining, visualization, and other capabilities.
Pentaho Big Data Overview
choose->prepare->visualize&explore

1 Choose Your Big Data Source

2 Auto-Prepare Data for Analysis

3 Interactively Visualize & Explore

http://www.pentaho.com/product/business-visualization-analytics#visual-analysis

Tatiana Yarmola, (ex) Intern UNECE Statistical Division.

12-Dec-2013
Platform Recommendations

• Processing Environment: **HortonWorks Hadoop distribution**

• Processing tools/software: **The Pentaho Business Analytics Suite Enterprise Edition**
  - Pentaho Business Analytics Suite Enterprise Edition provides a unified, interactive, and visual environment for data integration, data analysis, data mining, visualization, and other capabilities.
  - Pentaho's Data Integration component is fully compatible with Hortonworks Hadoop and allows 'drag and drop' development of MapReduce jobs.

• **R, RHadoop**, and possibly additional tools
Datasets Criteria

• Ease of locating and obtaining data from providers
• Cost of obtaining data (if any)
• Stability (or expected stability) over time
• Availability of data that can be used by several countries, or data whose format is at least broadly homogeneous across countries
• The existence of ID variables which enable the merging of big data sets with traditional statistical data sources

Tatiana Yarmola, (ex) Intern UNECE Statistical Division.
Datasets Specifications

• One or more from each of the categories below to be installed in the sandbox and experimented with for the creation of appropriate corresponding statistics:
  – Transactional sources (from banks/telecommunications providers/retail outlets)
  – Sensor data sources
  – Social network sources, image or video-based sources, other less-explored sources
Statistic Criteria

- conform to quality criteria for official statistics
- reliable and comparable across countries
- correspond in a systematic and predictable way with existing mainstream products, such as price statistics, household budget indicators, etc.

• Produce one or several
  1. statistics that correspond closely and in a predictable way with a mainstream statistic produced by most statistical organizations
  2. short term indicators of specific variables or cross-sectional statistics which permits the study of the detailed relationships between variables
  3. statistics that represent a new, non-traditional output
More details on Big Data Project and Sandbox Specifications at:

http://www1.unece.org/stat/platform/display/msis/Final+project+proposal%3A+The+Role+of+Big+Data+in+the+Modernisation+of+Statistical+Production

Thank You