Accessibility and data quality issues for selected indicators

Michael Nagy
11th Session of the Joint Task Force on environmental Indicators and Statistics
Geneva, 30 June – 1 July 2016

Content of the presentation

1. Accessibility of the national data
2. Why data quality and metadata matter?
3. Quality assurance in environment statistics
4. Conclusions and points for discussion
Access to data in different ways

- Interactive forms
- Tables on websites
- Downloadable xls-sheets

Cross-country comparison still difficult

- No central portal to access data for the same theme from different countries
- Language constraints
- We noticed that some weblinks have changed over time
- Data quality and data coverage is often unclear
- Often metadata is not available
Examples, why data quality and metadata matter – Example 1

Examples, why data quality and metadata matter – Example 2
Examples, why data quality and metadata matter – Example 3

Open questions:

a) What do empty cells mean? Zero or not applicable or no information available?

b) The renewable freshwater resources are equal to the internal flow. This can only be the case when the inflow equals zero.

Quality assurance in environment statistics

- Data have to be fit for purpose
- Environment Statistics uses a range of data sources, collected for various reasons, not always to official statistics standards
- Quality judgements are a result of holistic decisions based on
  - Uses
  - Costs
  - Conditions and circumstances affecting quality
  - User expectations
- Trade offs between accuracy and relevance

(Based on a presentation done by Adam Tipper, Statistics New Zealand)
The QA process in environment statistics follows usually 2 stages

1. Conceptually-focused quality assessment prior to data collection

2. Data checking: Ensuring that the actual data was compiled and prepared to the standard expected, and checking for missing values, outliers, unusual movements or levels completed once data has been received.

Data quality framework for environment statistics of Statistics New Zealand
Data quality criteria and their applicability to environment statistics (Statistics NZ)

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Standard descriptor</th>
<th>Applicability to environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevance</td>
<td>The degree to which the statistical product meets user needs in coverage, content and detail.</td>
<td>-Geographic coverage</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Fit to topic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Collection: How long, where from, by who and what for</td>
</tr>
<tr>
<td>Accuracy</td>
<td>The degree to which the information correctly describes the phenomena it was designed to measure.</td>
<td>-Accuracy in relation to topic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Methods and limitations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Available metadata</td>
</tr>
<tr>
<td>Timeliness</td>
<td>The degree to which data produced are up-to-date, published frequently and delivered to schedule.</td>
<td>Five years or less for key statistics</td>
</tr>
<tr>
<td>Accessibility</td>
<td>The ease with which users are able to access and understand the statistical data and its supporting information.</td>
<td>Extensive use of modelling</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Transparency</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Underlying data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Peer review</td>
</tr>
<tr>
<td>Coherence/consistency</td>
<td>The degree to which statistical information can be successfully brought together with other statistical information within a broad analytical framework and over time.</td>
<td>-Comparability with similar international indicators</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Coherency across measures</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Time-series consistency</td>
</tr>
<tr>
<td>Interpretability</td>
<td>The availability of supplementary information and metadata necessary to interpret and use the statistics effectively.</td>
<td>Ease by which a user can understand/track how the raw data feeds into the indicator.</td>
</tr>
</tbody>
</table>

Typical data quality problems with environmental data and statistics

- Not full coverage (e.g. only city data provided, but national data needed)
- Wrong unit of measurement (e.g. volume instead of mass)
- Decimal errors
- Samples are not representative (e.g. water emissions only measured when exceedance of a threshold value is most likely)
- Empty cells in a table. What does that mean (zero, not available or not applicable?)
- Outliers (measurement failure or not?)
- Use of different classifications (e.g. type of pollution classification for industries instead of ISIC)
- Same terms can have a different meaning by different expert communities (e.g. term “water use”)
- Not enough metadata available to assess the quality of data received from another source (e.g. administrative data)
What can NSOs do to improve data quality?

Before data collection (conceptual fit checks) clarify:

- Are the chosen data sources the right ones in terms of data quality, coverage, representativeness etc.?
- Can terms, definitions and classifications be matched with statistical standards?
- Which measurement units should be used?
- Is there a data quality assurance in place at the data source (e.g. for administrative data)?
- Which terminology needs to be used in questionnaires to avoid misunderstandings (example: the statistical term “water use” means something different for the water supply industry)
- Which is the right format for data exchange?
- Which metadata are needed?
- etc.

What can NSOs do to improve data quality?

After data collection carry out a data validation of the data:

- Coverage of data
- Correct use of measurement units
- Coherence of data with other sources
- Consistency of time series and outliers
- Balancing data items and sum of disaggregated data
- Correct use of classifications
- Metadata
- Etc.

- In case of data quality problems contact the original source of the data
- Some countries have thematic Intergovernmental Committees, which develop and quality assess jointly important national data sets (e.g. water, land use, air emissions etc.)
Conclusions
• Quality of environment statistics can be improved by measures at NSOs:
  – Before data collection
  – After data collection
• Staff involved needs a basic understanding of environment statistics concepts
  AND the frameworks that are used by the data sources (e.g. administrative
data sources) to “translate” correctly
• Metadata is important on all levels of the reporting chain
• A close involvement of data providers and/or feedback loops are recommended

Questions for the discussion
a) How should we address data quality issues in future meetings of the JTF?
b) What can be done to make country data sets easier accessible from outside the country and to directly compare data from different countries?
c) To which extent are metadata needed for the UNECE core indicators?
d) Do we need guidance on how to deal with “empty cells” (zero versus not applicable versus not available) for the UNECE indicators?