

Distr.
GENERAL

CES/AC.71/2005/13
4 March 2005

ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES
(EUROSTAT)**

**ORGANISATION FOR ECONOMIC
COOPERATION AND DEVELOPMENT
(OECD)
STATISTICS DIRECTORATE**

Joint ECE/Eurostat/OECD Meeting on the Management of Statistical Information Systems (MSIS)
(Bratislava, Slovakia, 18-20 April 2005)

Topic (ii): Development strategies for statistical information systems

**LINKING THE SDMX METADATA COMMON VOCABULARY
TO THE METADATA SYSTEMS OF INTERNATIONAL ORGANIZATIONS**

Invited Paper

Submitted by Eurostat and Organisation for Economic Cooperation and Development¹

I. SDMX AND THE METADATA COMMON VOCABULARY

1. The aims of the Statistical Data and Metadata Exchange (SDMX) initiative are, in general terms, the improvement in the efficiency in the exchange of statistical information (data and metadata) within the collective activities of the sponsoring organizations and the minimisation of the reporting burden of national reporting agencies. This can happen within the framework of a bilateral or multilateral exchange of information between parties, or with the placement of data and metadata on a location that can be accessed by all partners. In both cases, but particularly when the information is shared over the web, there is an essential need for users to understand the nature (and any limitations on use) of the statistics being exchanged. The SDMX standards aim to ensure that appropriate metadata always come along with the data: for this reason, standards for metadata exchange are extremely important in SDMX.

2. The Metadata Common Vocabulary (MCV for short) is one of the four main initiatives launched at the very beginning of SDMX². The aim is to develop a common understanding of standard cross-domain metadata components (or items) starting from the description of statistical concepts and methodologies used by statisticians in the collection, processing and dissemination of statistical data. The immediate objective was the development of a glossary of those standard components, whose definitions were consistent with any relevant existing international standards and guidelines, with the terminology used within SDMX organizations, within national agencies and, to the extent possible, in other related projects to develop international standards. The need for such work is evident if we only consider how many times the same metadata items are referred to by different names or, conversely, how many times the same name refers to different concepts.

¹ Prepared by Marco Pellegrino (Eurostat) and Denis Ward (OECD)

² Denis Ward and Marco Pellegrino, "Developing a common understanding of standard metadata components: a statistical glossary", Workshop on Statistical Data and Metadata Exchange, Washington, D.C., 6-7 September 2001.

3. The approach taken within MCV has been to focus on a system of definitions describing discrete metadata items (e.g. source, contact, periodicity, timeliness, reference period, coverage, or adjustment methods), which can be used for any statistical domain and independently from any general model or list of common metadata items developed by an organization. The Vocabulary is only concerned with the elaboration of these terminological building blocks, easily understandable and re-usable. The agreement on a common vocabulary would still provide the flexibility for each organization to manipulate metadata elements to derive a variety of specific formats and models according to their specific needs: an agreed-to list of generic metadata elements and associated definitions simply provides a common language. Each agency responsible for compiling metadata would then have the opportunity to organize metadata elements for deriving a variety of specific formats and models and for integrating data and metadata management.

II. POINT OF DEPARTURE AND CURRENT STATUS

4. The project built on work already undertaken by several organizations, rather than confusing the situation by the development of a whole new set of definitions. Where possible, definitions have been drawn from existing international standards or from statistical good practices. Where standard definitions were not available or not satisfactory, suitable national definitions have been considered or new definitions formulated.

5. The MCV covers a few types of items: general metadata terms (mostly derived from ISO/IEC 11179 and relevant UN-UN/ECE documents); metadata describing statistical methodologies (classifications, data collection, data editing, etc.), metadata for assessing quality and, quite obviously, a section of terms referring to data-metadata exchange and notably to the SDMX terminology (including GESMES terms). The value added is the opportunity of having one single entry point for accessing a variety of terms, sometimes not available or hard to find on the Internet. The MCV glossary is available on the web through extensive statistical glossary databases such as [CODED](#) (Eurostat concepts and definitions database, section "Metadata terminology") or the [OECD Glossary of statistical terms](#). Extractions are available in HTML/XML and in database format.

6. The present MCV draft (March 2005) consists of about 350 terms. It presents the following "fields": term, definition, source, related terms and context. The "context" field was used extensively throughout the glossary, sometimes for providing additional explanations, other times for highlighting peculiarities in how a certain definition is applied within a certain domain or geographical context.

7. The MCV is not intended to cover the whole range of statistical terminology, as this area is already covered by other general glossaries: the specific area of the MCV contains all those terms which are normally used for building and understanding metadata systems. A metadata glossary is necessarily linked to a series of other subject-specific glossaries (on classifications, on data editing, on subject-matter statistical areas) or to more universal statistical glossaries such as the CODED or OECD Glossary of Statistical Terms referred to above. These more extensive glossary databases also contain numerous terminologies and their definitions relevant to specific statistical domains (such as prices, national accounts, merchandise trade, etc.). The insertion within the MCV of some definitions derived from other glossaries should not be seen as a redundancy, but as a means of resolving the complex and interdisciplinary nature of metadata.

8. The MCV also contains an important standardisation aspect. In some cases, we have deliberately presented one definition and several context explanations for the same term, always quoting the respective source, in order to show their logical similarity or for highlighting a possible future work to make the existing definitions converge. This is the case for some "quality" items (accuracy, accessibility, etc.): users can live with different quality frameworks or different meta-models, as long as each item is well identified, defined and known by users. In other words, transparency is a pre-requisite for a correct interpretation (and for convergence) of any statistical framework.

III. THE USE OF THE MCV

8. The MCV was initiated as a joint Eurostat-OECD effort. Work is now proceeding, together with the other SDMX partners, for updating existing definitions, for ensuring that only relevant items are defined and for promoting a more active use of this tool.

9. Within Eurostat and the OECD, the MCV is already used to ensure clarity and terminological consistency for the respective metadata frameworks used in their corporate metadata repositories (Eurostat free dissemination, OECD Metastore). The use of standard definitions taken from the MCV is encouraged within each metadata-related project and activity.

10. The availability of a web repository of standard definitions, available for all Internet users, should also be regarded as a unique chance for creating a common understanding of metadata terms. CODED provides a unique reference point for Eurostat and EU members, and the same thing can be said for the OECD glossary with reference to the OECD area.

11. About the use of MCV standard definitions within the SDMX initiative, the 2005 work programme includes the release of SDMX version 2 which puts more emphasis on the identification of reference metadata³ items and subject-matter domains where to test metadata standards and advanced technical standards for data/metadata exchange. The harmonization of statistical metadata entails not only the definitions of the concepts and their names, but also, where appropriate, their representation with standard code lists and the role they play within key family structures for data exchange. In this context, the Metadata Common Vocabulary plays an important role in providing the common set of terms and definitions that can be used to describe the data.

12. Agreement on a content standard for SDMX common metadata concepts implies a dynamic update of the MCV to reflect the SDMX standards. In many instances, the “context” field of the MCV would be updated to define the SDMX application of the term. In other cases, when the metadata concepts and the granular items falling under the core set of terms will be revised, the MCV will need to include new terms and refine existing definitions.

IV. FEEDBACK REQUIRED

13. The draft MCV (available at www.sdmx.org) is open to external contributions, criticisms and suggestions: the MCV intends to take advantage of any published or unpublished glossary and from research initiatives.

14. Feedback and suggestions are sought by the project managers - Marco Pellegrino (marco.pellegrino@cec.eu.int) and Denis Ward (denis.ward@oecd.org) - particularly in the following areas:

- Relevance of terms included in the Vocabulary;
- Suggestions for additional metadata items;
- Availability of more appropriate definitions;
- Use of MCV in connection with metadata management systems.

³ SDMX metadata standards build on the distinction between “structural” and “reference” metadata. *Structural metadata* are metadata acting as identifiers and descriptors of the data, such as names of variables or dimensions of statistical cubes. *Reference metadata* are metadata describing the contents and the quality of the statistical data: conceptual metadata (describing the concepts used and their practical implementation), methodological metadata (describing methods used for the generation of the data) and quality metadata (describing the different quality dimensions of the resulting statistics, such as timeliness and accuracy). Reference metadata are sometimes produced, collected or disseminated separately from the statistical data to which they refer.

ANNEX**LIST OF TERMS DEFINED WITHIN THE METADATA COMMON VOCABULARY (MARCH 2005)**

- | | | | |
|-----|--|------|-----------------------------------|
| 1. | Accessibility | 56. | Coverage ratio |
| 2. | Accounting basis | 57. | Cut-off survey |
| 3. | Accuracy | 58. | Cut-off threshold |
| 4. | Activity | 59. | Data |
| 5. | Adjustment methods | 60. | Data analysis |
| 6. | Administered item (ISO) | 61. | Data capture |
| 7. | Administration record (ISO) | 62. | Data checking |
| 8. | Administrative data | 63. | Data collection |
| 9. | Administrative source | 64. | Data collection, administrative |
| 10. | Aggregation | 65. | Data collection, survey |
| 11. | Analytical framework | 66. | Data confrontation |
| 12. | Area sampling | 67. | Data Dissemination Standards, IMF |
| 13. | Attachment level | 68. | Data editing |
| 14. | Attribute (ISO) | 69. | Data editing, graphical |
| 15. | Attribute (GESMES) | 70. | Data element concept (ISO) |
| 16. | Base period | 71. | Data element derivation (ISO) |
| 17. | Base weight | 72. | Data element, derived |
| 18. | Basic attribute (ISO) | 73. | Data element (ISO) |
| 19. | Basic statistical data | 74. | Data exchange |
| 20. | Benchmark | 75. | Data exchange context (Gesmes/TS) |
| 21. | Benchmarking | 76. | Data identifier (ISO) |
| 22. | Bias | 77. | Data interchange |
| 23. | Census | 78. | Data item (ISO) |
| 24. | Chain Index | 79. | Data item (UN) |
| 25. | Changes in classifications and structure | 80. | Data model (ISO) |
| 26. | Characteristic (ISO) | 81. | Data processing |
| 27. | Clarity | 82. | Data provider |
| 28. | Class (ISO) | 83. | Data reconciliation |
| 29. | Classification | 84. | Data set |
| 30. | Classification scheme (ISO) | 85. | Data source |
| 31. | Classification, standard | 86. | Data status (upon release) |
| 32. | Classification unit | 87. | Datatype (ISO) |
| 33. | Code list (Gesmes/TS) | 88. | Date |
| 34. | Coding | 89. | Date, creation (ISO) |
| 35. | Coding error | 90. | Date, effective (ISO) |
| 36. | Coefficient of variation | 91. | Date of last change (ISO) |
| 37. | Coherence | 92. | Definition |
| 38. | Comparability | 93. | Definition, preferred (ISO) |
| 39. | Compilation practices | 94. | Definition, structural |
| 40. | Completeness | 95. | Derivation input (ISO) |
| 41. | Computation of lowest level indices | 96. | Derivation output (ISO) |
| 42. | Computer-Assisted Interviewing, CAI | 97. | Derivation rule (ISO) |
| 43. | Concept (ISO) | 98. | Derived statistic |
| 44. | Conceptual data model (ISO) | 99. | Dimension (GESMES/TS) |
| 45. | Conceptual domain (ISO) | 100. | Dimensionality (ISO) |
| 46. | Confidential data | 101. | Disaggregation |
| 47. | Confidentiality | 102. | Disclosure analysis |
| 48. | Consistency | 103. | Dissemination |
| 49. | Consolidation (national accounts) | 104. | Dissemination format |
| 50. | Contact (ISO) | 105. | Documentation |
| 51. | Context (ISO) | 106. | Domain |
| 52. | Co-ordination of samples | 107. | Dublin Core |
| 53. | Country identifier (ISO) | 108. | EDIFACT (ISO) |
| 54. | Coverage | 109. | Electronic data interchange (EDI) |
| 55. | Coverage errors | 110. | Entity (ISO) |

111. Entry, terminological (ISO)
112. Error of estimation
113. Error of observation
114. Error, statistical
115. Estimate
116. Estimation
117. Estimator
118. Expected value
119. Flag
120. Flow series/data
121. Follow-up
122. Footnote
123. Frame
124. Frame error
125. Frequency of time series
126. Gateway
127. GDDS
128. General Data Dissemination System (GDDS)
129. GESMES
130. GESMES/CB
131. GESMES/TS
132. GESMES/TS data model
133. Glossary
134. Grossing/Netting
135. Hierarchy
136. Identifier (ISO)
137. Imputation
138. Index number
139. Indicator, statistical
140. Industry
141. Information
142. Information system
143. Inlier
144. Institutional framework
145. Integrity
146. Internal access
147. International code designator (ISO)
148. Interpolation
149. Interviewer error
150. ISO/IEC 11179
151. Item
152. Item response rate
153. Key family
154. Key structure
155. Key (time series or sibling group)
156. Keyword (ISO)
157. Language (ISO)
158. Levels of data
159. Longitudinal data
160. Macrodata, statistical
161. Macro-editing
162. Maintenance agency (Gesmes/TS structural definitions)
163. Mean-square error
164. Measurement error
165. Metadata dimension (SDDS)
166. Metadata (ISO)
167. Metadata item (ISO)
168. Metadata layer
169. Metadata object (ISO)
170. Metadata registry (ISO)
171. Metadata set (ISO)
172. Metadata, statistical
173. Metadata system, statistical
174. Metainformation, statistical
175. Metainformation system, statistical
176. Metamodel (ISO)
177. Methodological soundness
178. Methodology
179. Methodology, statistical
180. Microdata, statistical
181. Micro-editing
182. Ministerial commentary
183. Misclassification
184. Missing data
185. Model assumption error
186. Name (ISO)
187. Nomenclature
188. Non-probability sample
189. Non-response
190. Non-response bias
191. Non-response error
192. Non-response rate
193. Non-sampling error
194. Not seasonally adjusted
195. Number raised estimation
196. Object class (ISO)
197. Object class term (ISO)
198. Object (ISO)
199. Objectives
200. Observation
201. Observation, pre-break
202. Observation unit
203. Ontology
204. Organization (ISO)
205. Organization identifier (ISO)
206. Organization, responsible (ISO)
207. Outliers
208. Out-of-scope units
209. Over-coverage
210. Period
211. Periodicity
212. Permissible value (ISO)
213. Permitted value (ISO)
214. Population, statistical
215. Precision
216. Prices, types of
217. Primary data
218. Primary source (of statistical data)
219. Probability sample
220. Processing error
221. Product
222. Property (ISO)
223. Provider load
224. Public disclosure
225. Punctuality
226. Qualitative data
227. Quality control survey
228. Quality differences
229. Quality (Eurostat)
230. Quality (IMF)
231. Quality index
232. Quality (ISO)
233. Quality (OECD)

- 234. Quality, prerequisites of
- 235. Quantitative data
- 236. Questionnaire
- 237. Questionnaire design
- 238. Ratio estimation
- 239. Recommended use of data
- 240. Record check
- 241. Recording of transactions
- 242. Record-keeping error
- 243. Reference document (ISO)
- 244. Reference metadata
- 245. Reference period
- 246. Reference time
- 247. Refusal rate
- 248. Register
- 249. Registrar (ISO)
- 250. Registration authority (ISO)
- 251. Registration (ISO)
- 252. Registry item (ISO)
- 253. Registry metamodel (ISO)
- 254. Related data reference (ISO)
- 255. Related metadata reference (ISO)
- 256. Relationship (ISO)
- 257. Release calendar
- 258. Relevance
- 259. Reliability
- 260. Reporting unit
- 261. Respondent burden
- 262. Respondent load
- 263. Response errors
- 264. Response rate
- 265. Revision policy
- 266. Sample
- 267. Sample design
- 268. Sample size
- 269. Sample survey
- 270. Sampling
- 271. Sampling error
- 272. Sampling fraction
- 273. Sampling frame
- 274. Sampling techniques
- 275. Sampling unit
- 276. Schedule
- 277. Scope
- 278. SDDS
- 279. SDMX
- 280. Seasonal adjustment
- 281. Secondary sources (of statistical data)
- 282. Sector, institutional
- 283. Security, data
- 284. Semantics (ISO)
- 285. Serviceability
- 286. Sibling group
- 287. Simultaneous release
- 288. Source
- 289. Special Data Dissemination Standard (SDDS)
- 290. Special language (ISO)
- 291. Standard error
- 292. Standard error, relative
- 293. Statistical concept
- 294. Statistical Data and Metadata Exchange (SDMX)
- 295. Statistical measure
- 296. Statistical message
- 297. Statistical metadata repository
- 298. Statistical production
- 299. Statistical standard
- 300. Statistical standard, international
- 301. Stewardship (ISO)
- 302. Stock series/data
- 303. Structural metadata
- 304. Stratification
- 305. Submission (ISO)
- 306. Submitting organization (ISO)
- 307. Survey
- 308. Survey design
- 309. Syntax (ISO)
- 310. Target population
- 311. Taxonomy (ISO)
- 312. Term
- 313. Terminological system (ISO)
- 314. Terminology
- 315. Thesaurus (ISO)
- 316. Time of recording
- 317. Time series
- 318. Time series breaks
- 319. Timeliness
- 320. Trend
- 321. Trend estimates
- 322. True value
- 323. Type of data collection
- 324. Under-coverage
- 325. Unit, analytical
- 326. Unit, institutional
- 327. Unit non-response
- 328. Unit of measure (ISO)
- 329. Unit response rate
- 330. Unit, statistical
- 331. Unit value
- 332. Unit value index
- 333. User needs (for statistics)
- 334. User satisfaction survey
- 335. Validation
- 336. Valuation
- 337. Value domain (ISO)
- 338. Value item (ISO)
- 339. Value meaning (ISO)
- 340. Variable
- 341. Variance
- 342. Variance estimation
- 343. Verification
- 344. Weight
- 345. XML
- 346. Year-to-date data