

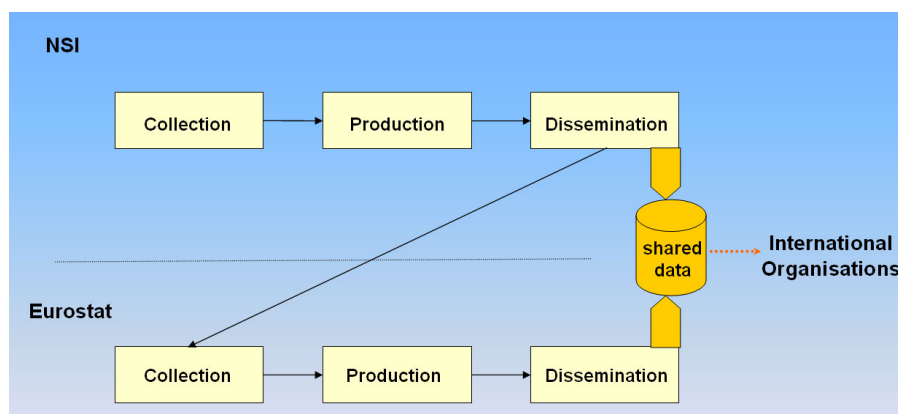


Session 2: Coherence of data published by various international organizations

Using SDMX standards for dissemination of short-term indicators on the European economy ¹

I. BACKGROUND INFORMATION

1. The SODI project focuses on the interoperability of statistics for collecting and disseminating short-term statistics, especially in the domains of the Principal European Economic Indicators (PEEI), with the overall objective of increasing timeliness and accessibility. SODI (SDMX Open Data Interchange) is an SDMX implementation project, which means that it is one of the official proofs of concept for SDMX. It is comparable in scope to the "National Accounts World Wide Exchange" project (NAWWE) undertaken by the OECD.
2. The main benefits expected from SODI are:
 - improved quality and timeliness of statistics;
 - reduced reporting burden, through the use of common formats for data exchange and data sharing of statistical information on web sites to complement or replace direct reporting;
 - more user-friendly access to data and related metadata, for business users and citizens, when publishing international and national statistics on the web;
 - reduced human resources needed to process the data in the Competent National Authorities (CNA) and Eurostat.
3. SODI is a data sharing project in the European Statistical System. The Data Sharing Model is a mechanism by which data (e.g. statistics) are made available to users in a common environment (the Internet) in a common technical format and with agreed common codes and metadata.



¹ Prepared by B. A. Lindblad, L. Maqua, M. Pellegrino and G. Sindoni - Eurostat. This paper has been submitted to the METIS Work session on statistical metadata, held in Geneva on 3-4-5 April 2006.

4. In this model, users locate and retrieve the data relevant to their needs using a registry made available to them by those partners participating in the data sharing exercise. The model, currently being tested for the dissemination of the Principal European Economic Indicators, is being tested on other domains. In the SODI project, the shared data are maintained by Eurostat.

5. Before SODI was launched, a pilot project has been conducted. A Task Force with Eurostat and the national statistical institutes of Germany, France, the Netherlands, Sweden and the UK performed trials, testing different ways of transmission (Push – the traditional method of sending data to Eurostat, and Pull – the data sharing approach via web services) and different data formats (SDMX-ML – the XML version of the SDMX data format, and SDMX-EDI, the GESMES-TS compatible version of SDMX). The results of the trials were successful:

Country	Format	Method	Status
DE	SDMX-ML	Pull	successful transmission
FR	SDMX-EDI	Push	successful transmission
NL	SDMX-ML	Pull	successful transmission
SE	SDMX-EDI	Push	successful transmission
UK	SDMX-ML	Push	successful transmission

6. The SODI pilots had two main deliverables: a report on the issues encountered, which was delivered to the FROCH² group in June 2005, and a proof of concept for the data sharing approach, which was delivered, together with live demonstrations, in November 2005 at the FROCH group and the SPC³.

7. The conclusions of the SODI pilots exercise have been drawn as follows:

- The approach of SODI, based on SDMX standards, is technically feasible: work should continue towards the objective of opening SODI to public access in the second half of 2006.
- The SODI pilots have enabled the identification of the issues that must be taken in order to pass to an operational implementation of the SODI concept in terms of widening the range of indicators. In principle, SODI aims to cover all the PEEI. However, experience has shown that for the data-sharing approach to work, certain criteria concerning the choice of indicators have to be fulfilled, such as the availability of a key family fully compatible with the rules used in GESMES/TS and SDMX-ML; an acceptable level of data quality; and an acceptable level of harmonisation in the national versions of the indicators.
- The SODI pilots have enabled the identification of the steps that must be taken into account for covering more countries. The pilots have enabled countries to identify the nature of the work required and hence to determine whether costs are manageable. In general it appears that this should be the case.
- The SODI TF should continue work as its input will be needed to deal with several of the issues identified in the Issue Report.

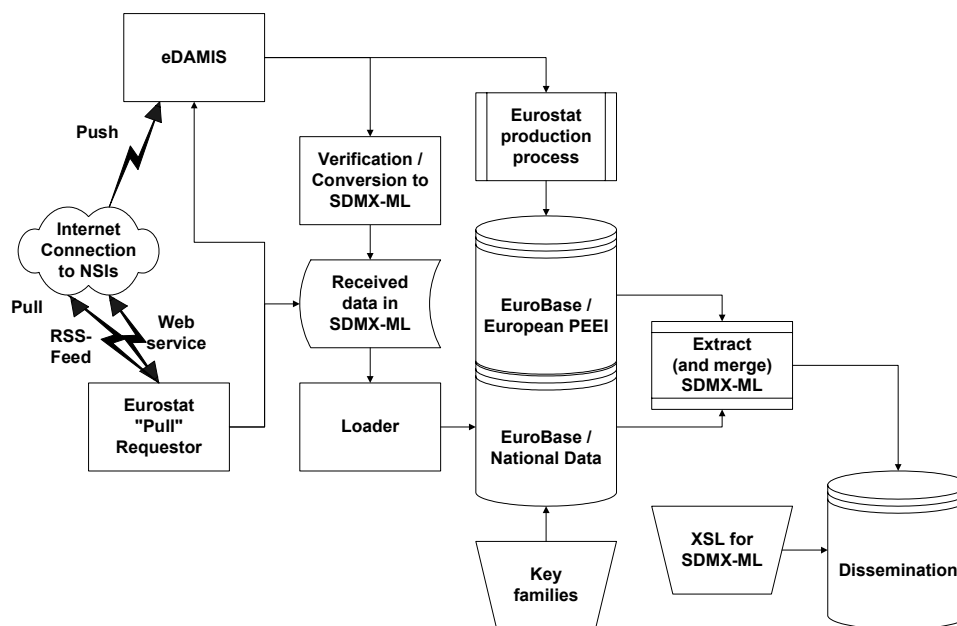
² Friends of the Chair, a high level group which acts as a think-tank to the Statistical Programme Committee

³ Statistical Programme Committee, comprising the presidents of the National Statistical Institutes of the European Statistical System.

II. THE CONCEPTS AND IMPLEMENTATION OF THE SODI PROJECT

A. The SODI Approach towards Data Sharing and SDMX Implementation

8. The SODI process accepts both SDMX-EDI and SDMX-ML as an input, and can receive data both via eDAMIS⁴ and from a web service set up by the competent national authority on its web site. In the further processing up to dissemination, SODI only uses the SDMX-ML format. So, despite being some type of shortcut to Eurostat's normal data processing cycle, SODI fits into the data life cycle of Eurostat. The processing of data in SODI is synthesised by the following picture:



B. Project Organisation and Budget

9. The project is financed as an action of the X-DIS project (XML for Data Interoperability in Statistics) by the Commission programme IDABC (Interoperable Delivery of European eGovernment Services to public Administrations, Businesses and Citizens) for the budget years 2005 to 2008.

10. The project requires a close co-operation with the Member States, which is coordinated by the SODI task force described in the following paragraph.

C. The SODI Task Force

11. The SODI task force has been enlarged since the pilots. Now, the National Statistical Institutes of the following countries participate in the task force (with ECB and OECD as observers): Denmark, Germany, France, Italy, the Netherlands, Norway, Slovenia, Sweden and the UK.

12. The members of the SODI task force

- support Eurostat in the SODI implementation;
- are consulted on the SODI work plan and other important documents on SODI and SDMX;

⁴ eDAMIS is a system aimed at implementing Eurostat's concept of a "single entry point" for statistical data. That is the hub where data sets should be sent by national competent authorities and delivered to competent Eurostat's production units.

- give advice on SODI issues;
- send data to Eurostat to be used in the SODI process;
- receive technical support by Eurostat and its consultants on SDMX and the implementation of SODI.

III. ISSUES TACKLED BY SODI

13. This paragraph summarises the issues encountered or identified during the SODI pilots. It marks points for decision, threats and opportunities, and gives recommendations and lessons learned, tackling both technical and non-technical issues.

A. Technical issues

SDMX-EDI and SDMX-ML

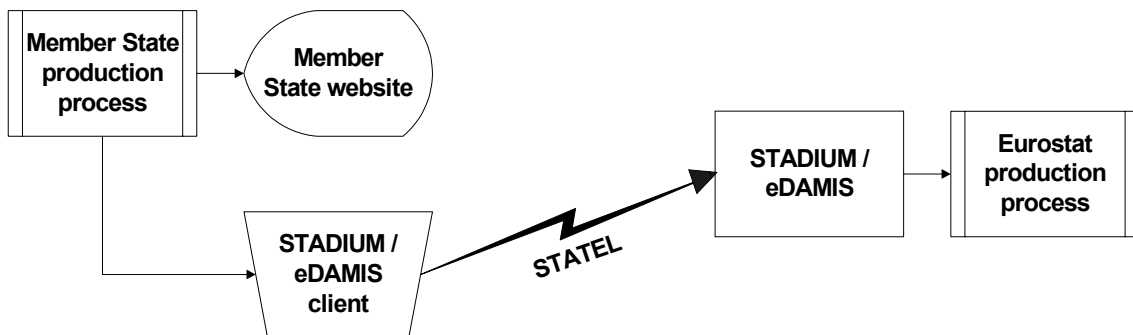
14. One of the main results of the pilots is that the existence of two different data formats does not cause any real problem. The conversion from SDMX-EDI to SDMX-ML within the pilots is being done with a simple "home-made" tool, which shall be replaced in the future by a more appropriate one, based on open source software and including also structural definition maintenance and conversion to other formats.

SDMX-ML for dissemination

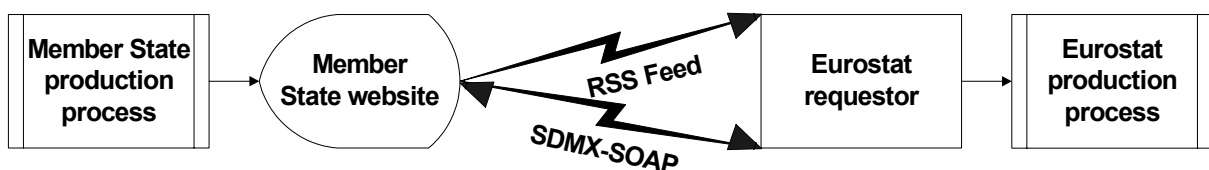
15. The SODI project will test as well the use of SDMX-ML for dissemination. For this, formatting information (in XSL, the eXtensible Stylesheet Language, or CSS, Cascading Style Sheets) will be added to the SDMX-ML data, so that they can be visualised in any modern browser.

Push and Pull

16. The Push method in SODI means that the Member States send their data to Eurostat using the Single Entry Point (SEP) currently implemented using a system named STADIUM or the new eDAMIS system:



17. The Pull method is characterised by the fact, that a Eurostat application (called "requestor") fetches the data from the web site of a National Authority:



18. The requestor is triggered by an RSS feed. RSS stands for “Really Simple Syndication”, and is the web standard for news feeds. The implementation of the Pull method is currently under construction in both Eurostat and two National Statistical Institutes (Netherlands and Germany).

19. The Pull method is more difficult to implement both at Eurostat and at the NSIs; because of this, it has been the subject of in-depth technical investigations within the SODI pilots. However, this method will finally fit better into a general SDMX-ML dissemination environment in the Member States, so that a seamless integration into the production process is automatically achieved. In addition, the Pull uses only standardised internet compatible methods. The Pull approach guarantees that data are available at the same time for Eurostat as they are published nationally, while it requires that the National Statistical Institutes publish data conforming to the European concepts.

B. Standardisation issues

Statistical terminology

20. The development of more efficient processes for sharing data requires the adoption of a standard terminology for describing the statistics being exchanged. The Metadata Common Vocabulary (MCV) elaborated under the SDMX initiative provides the common set of terms (and related definitions) to be used for the sake of terminological consistency. Agreement on such a standard implies a continuous update to reflect core concepts used within SDMX and with national institutes. SODI, therefore, is one of those initiatives which can provide a valuable “reality check”, through the description of data structures and the attachment of a set of reference metadata documenting the data. Existing ambiguities in the use of a term, or the fact that not all terms have been identified in the MCV yet, call for a parallel expansion of the MCV during 2006.

Structural metadata for “Data Structure Definitions” (Key Families)

21. It is a prerequisite for the full SODI implementation that all key families are SDMX-compliant. This was not the case for the existing structural definition on GDP, which was not compatible with GESMES/TS (and hence with SDMX). GDP data are collected in the European System of Accounts (ESA95) framework. The ESA95 structural definition was recently revised using a SDMX-compliant standard (GESMES/TS which can be regarded as equivalent to SDMX-EDI). A thorough analysis of existing GESMES structural definitions of the PEEI, currently being performed, will allow to define priority domains for SDMX implementation and requirements to migrate from GESMES to SDMX.

Reference metadata

22. In accordance with the general principle that no data should be made available without an acceptable coverage of metadata, Eurostat conducted a review of metadata available from national, European or international web sites. The main issues were: a) the availability of sufficient metadata coverage in multiple languages; b) the conceptual overlap between formats used within national and international web sites; c) issues concerning the updating and dissemination process of the different metadata items.

23. The metadata coverage associated to the first test domains is quite good, although the same kind of information is available from different providers (national web sites, Eurostat, IMF) and not all of the metadata are still available in English from all countries. The need of advancing towards a more coordinated framework which links national and EU metadata has been stressed by several participants in the SODI task-force.

24. Eurostat, in coordination with member States, intends to make use of the latest standards for associating data with a consistent and standardised set of metadata items. In this context, the use of SDMX standards would allow the harmonisation of presentation styles and at the same time the standardisation of data and metadata descriptions, so that these can be exchanged, read and processed by computers without manual intervention.

25. Through the use of standard concepts, applied to the exchange of data sets (on the basis of the formal definition of the data structure) as well as to metadata sets (on the basis of the definition of the

metadata structure), there is the concrete possibility of setting the requirements for a European concept family of reference metadata to be exchanged and shared by using web services to navigate, find and process the information.

26. The implementation of such a system implies that institutions preparing metadata in a standard format on their websites will make it possible for Eurostat (and other organisations) to access this information from the web rather than putting in place ad hoc transmissions in various formats. The progress made on the identification of commonalities in the existing metadata systems and on the standardisation of the terminology and concepts used (see the SDMX work on cross-domain concepts and on the “Metadata Common Vocabulary”) will help reducing the metadata reporting burden of national institutes and, at the same time, will improve the quality and consistency of metadata descriptions across countries.

27. While working on the technical infrastructure, Eurostat is currently improving the granularity of the reference metadata format, with the aim of extending the conceptual coverage of the format and in particular for incorporating more elements on quality assessment, according to the criteria identified by the European statistical code of practice. The modular list of concepts (described in Annex 2) is built on the current format used within Eurostat, with some limited extensions on quality elements which are going to be further detailed by the end of this year. The current list is going to be used for testing the possibility of disseminating a good selection of reference metadata with regard to PEEI data.

C. Statistical issues

Concepts

28. As mentioned in the previous paragraph, the harmonisation of concepts is indispensable for sharing data. Fortunately, for most PEEI and National accounts data (ESA95) this is already the case. For other data flows, for instance in the area of social statistics, this has to be checked case by case before integrating them into SODI. Especially when the Pull method is used, it is indispensable that Member States and Eurostat use the same concepts for publishing data. This also includes coordination on the possibility of disseminating seasonally adjusted figures for infra-annual data, on the method used and on the provision of relevant methodological explanations to the user.

29. Eurostat, under the “Data Life Cycle initiative” (CVD, Cycle de Vie des Données) plans to reduce the number of code lists in use for the same concept. SODI – by requiring a unique concept for reception, production and dissemination – contributes to this effort.

Statistical confidentiality

30. At the moment, we are not planning to cover data flows where all or part of the data are confidential. However, this has to be checked for all data flows before they are considered by SODI; and in case of (partial) confidentiality this has to be treated correctly by the dissemination modules.

Validation

31. SODI assumes that national data are for publication. However, even when Eurostat is taking the national value “as it is”, a minimum amount of technical verification has to be done to prevent from human errors or technical problems in the transmission process leading to application failure or inconsistent data. So, at least a formal verification (correctness of XML syntax, adherence to SDMX standard, compliance with code lists) has to be done. More statistical validation should only be performed, at this stage, when fully automated. In addition, the response to an erroneous message has to be defined.

Footnote Treatment

32. Footnotes are part of the SDMX data model, so the processing of footnotes in any environment conforming to SDMX is neither a problem for the standard nor a technical issue. However, in addition to footnotes received with the SDMX message containing the data, there may be additional footnotes

at different levels (footnotes might apply to a single observation, a country, a period or some other dimension of the data) which have to be treated correctly. First, in the output national data have to be correctly tagged as “national”; second, footnotes added by a Eurostat production unit have to be applied correctly; third, there may be standard footnotes for certain dimensions, which are defined by Eurostat, but have to be applied as well (or exclusively) to national data. This could be the case when data do not include the whole country (for instance, German data before October 1990) or a country uses a slightly different concept for one dimension (e.g. a different method of seasonal adjustment).

D. Political Issues

The role of Eurostat in a data sharing environment

33. In a data sharing environment, the role of Eurostat has to be redefined. Will Eurostat only become a coordination body, responsible for the harmonisation of concepts and methods, and maybe with a stronger role in quality assessment, or will Eurostat do more than just compile the national data? Which data treatment will still remain to Eurostat? How to manage shared responsibility with other organisations with respect to the maintenance and update of statistical structures (structural definitions, code lists, etc.)?

Aggregates

34. Especially delicate is the question of aggregates (like EU25, EU15, Euro-zone), which are calculated by Eurostat. Will these aggregates only be calculated for data treated by Eurostat, or will it just be the aggregate of the data available. It might have to be explained a situation where a European aggregate does not correspond to the data in a given table.

Releases

35. This regards the treatment of different releases of the same data: at the moment, the basic idea is to mark, in the Eurostat tables, the incoming data as “national”, before they have been processed by Eurostat and replaced by “European” data. This becomes a problem in the (frequent) case of successive releases. When releases come, will the “national” updates substitute the (older) “European” ones, or will they stay invisible, until they have been processed by Eurostat? And what about the aggregates in this case (there might even be a different aggregate policy for original and updated data).

Embargo Policies

36. Short term statistical data are often published under an embargo policy, i.e., data may not be published before a certain date and time, however, this often differs in the European Statistical System (especially concerning dates and handling of delays). It has to be clarified whether SODI will require a harmonisation. A special question arises when Eurostat’s embargo date is later than the national ones. Should in this case the “national” data already be published, of course without a European aggregate?

Ownership of the Data

37. At the moment, it is planned that Eurostat marks the incoming data in a footnote as “national” and removes this footnote after validation by the production unit. So Eurostat would distinguish between data, for which it takes the ownership and responsibility, and national data, where the National Statistical Institutes are the responsible owners.

E. Organisational Issues

Coexistence of Standards and Methods

38. As Member States might progress with different speeds or set different priorities for their dissemination systems, we might experience a very long period with both Push and Pull method and both SDMX-ML and SDMX-EDI used in parallel. Eurostat’s data reception environment has to care for an automatic transparent integration and conversion process, unless there is an explicit expressed will in the ESS to agree on a single method and/or a single data format.

Embargo Treatment

39. Currently, embargos are handled by the production unit processing the data. In the future, if Eurostat ceases the manual treatment of data for certain data flows, the Eurostat embargos have to be handled by the dissemination environment, while the national embargos have to be handled by the reception environment (this applies only for the “Push” approach, with “Pull” we expect data not be available to Eurostat before the national publication). A special case arises when the national and the European data shall be published simultaneously. In this case, the “Pull” as currently designed will not achieve precisely synchronous publication, as the necessary processing by Eurostat will cause a delay for the publication of the European data. It might be necessary to redefine slightly the objective of “simultaneous publication”.

Integration of SODI into Eurostat’s Data Life Cycle

40. In its basic idea, the publication of national data on Eurostat’s web site is opposite to the data life cycle project of Eurostat, as neither the reception nor the production environment is used in the case of the “Pull” model. On the other hand, the idea of a single entry point and a reduction of the number of production systems are vital for the correct functioning of Eurostat’s IT. So, the Pull method will be integrated into the single entry point. In particular, the requestor will be integrated into the eDAMIS system. In addition, eDAMIS will learn to handle SDMX-ML as a generic format (like it handles GESMES today) not requiring a separate envelope for the metadata. In the production environment, a special “SODI” process will be created. Although this is at the moment an additional process, finally, with further harmonisation in the ESS, several currently different processes could be given up in sake of this single process.

F. Legal Issues

“Pull” and the Obligations of Member States

41. Before using the Pull method in production for data flows covered by Commission or Council Regulations, it has to be clarified if this method fulfils any obligations of the Member States to deliver data to Eurostat. Legally, there can be a difference between the obligation to deliver data (without being explicitly asked) or to provide the data on request (as in the Pull method).

SODI and SDMX in Legal Acts

42. Normally, a general reference to a standardised format in legal acts is preferred, rather than a specification of the data format. This is wise as the lifecycle of legal acts is normally longer than the lifecycle of data transmission methods. Although we expect XML formats like SDMX-ML to have a very long lifetime, technical progress on transmission protocols or XML related standards (like IPv6, XML security, standardisation of web services, and so on) will influence the SDMX standard and its implementations. On the other hand, we expect SDMX-ML – with the standardisation by ISO and the commitment of the stakeholders – to become a widely accepted standard, so it will be perfectly covered by most existing legal acts.

Principal European Economic Indicators List

Set 1: Price Indicators

- 1.1. Harmonised Consumer Price Index: MUICP flash estimate: release end of reference month
- 1.2. Harmonised Consumer Price Index: actual indices: release 2,5 weeks after reference month

Set 2: National Accounts Indicators

- 2.1. Quarterly National Accounts: flash GDP: release t+45
- 2.2. Quarterly National Accounts: first GDP release with breakdowns: t+60
- 2.3. Quarterly National Accounts: Sector Accounts: release t+90
- 2.4. Quarterly Government Finance Statistics: release t+90

Set 3: Business Indicators

- 3.1 Industrial production index: release t+30
- 3.2 Industrial output price index for domestic markets: release t+35
- 3.3 Industrial new orders index: release t+50
- 3.4 Industrial import price index: release t+30
- 3.5 1. Production in construction: quarterly: release t+45
2. Monthly: release t+30
- 3.6 Turnover index for retail trade and repair: release t+30
- 3.7 Turnover index for other services: release t+60
- 3.8 Corporate output price index for services: release t+60

Set 4: Labour Market Indicators

- 4.1. Unemployment rate: release t+30
- 4.2. 1: Job vacancy rate: quarterly
2: monthly: release t+30
- 4.3. 1. Employment: monthly release t+30
2. quarterly: release t+45
- 4.4. Labour cost index (US: Employment cost index) release t+60

Set 5: Foreign Trade Indicators

- 5.1. External trade balance:
intra- and extra-MU; intra- and extra-EU: release t+46

EUROSTAT – SDMX CROSS-DOMAIN CONCEPTS MAPPING

EUROSTAT DISSEMINATION METADATA CONCEPTS		MAPPING TO CURRENT DRAFT OF SDMX CROSS-DOMAIN CONCEPTS
Top level	Child level	
Metadata Update	Last certified without update	Date of update
	Last update of content	Date of update
Contact	Organisation	Contact
	Address	Contact
	Contact name or service	Contact
	e-mail address	Contact
Data coverage	Short description of data domain	Data presentation
	Data breakdown and main variables	Data presentation
	Units of measure	Data presentation
Periodicity	Periodicity of compilation	Frequency and periodicity
	Database frequency	Frequency and periodicity
Timeliness and punctuality	Timeliness	Timeliness and punctuality
	Punctuality	Timeliness and punctuality
Transparency of practices	Legal acts, reporting requirements	Institutional framework
	Rules on confidentiality	Institutional framework
	Internal access	Transparency
	Commentary on the occasion of release	Transparency
Accessibility	Notification of changes in methodology	Transparency
	Release calendar	Release calendar
	Simultaneous release	Simultaneous release
	Dissemination formats	Dissemination formats
Quality cross-checks	Documentation on methodology	Accessibility of documentation
	Related data and quality cross-checks	[No direct concordance]
Accuracy and reliability	References to quality reports	[No direct concordance]
	Overall accuracy assessment	Accuracy
Comparability and coherence	Quality checks before release	Accuracy
	Comparability over time	Comparability and coherence
	Comparability over space	Comparability and coherence
	Comparability with related sources	Comparability and coherence
	Comparability between datasets	Comparability and coherence
Relevance	Breaks in time series	Comparability and coherence
	Rate of available statistics (user needs)	Relevance
	Intended audience and purpose	Relevance
	Supplementary data	Supplementary data
Statistical concepts and classifications		
	Statistical concept	Statistical concept
	Definition of indicators	Statistical concept
	Classification system	Classification systems
	Conformity with official standards	Classification systems
Scope of the data	Classification coverage	Classification systems
	Reference area / geopolitical entity	Scope/coverage
	Time coverage	Scope/coverage
	Statistical unit	Scope/coverage
Accounting conventions	Statistical population	Scope/coverage
	Reference period	Accounting conventions
	Base period	Accounting conventions
Nature of basic data	Basis for recording	Accounting conventions
	Data source used	Source data
	Type of survey	Source data
Compilation practices	Methods of data collection	Source data
	Compilation	Statistical processing
	Adjustments and weights	Statistical processing
	Data validation	Statistical processing
Other	Revision policy and practice	Revision policy and practice
	Warnings on re-use and limitations	[No direct concordance]