SYRTO

SYstemic Risk TOmography
Signals, Measurements, Transmission Channels, and Policy Interventions

SYRTO project – data quality framework

Chiara Carini

Technical report

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n° 320270.

This documents reflects only the author's view. The European Union is not liable for any use that may be made of the information contained therein.

SYRTO project – data and quality assessment framework

Chiara Carini[1]

In defining and implementing a statistical process, one of the key steps is quality assessment. Indeed, it is now widely accepted that no systematic evaluation of data quality can result in the loss of one or more steps of a statistical process (UN Statistics Division, 2003; Eurostat, 2009).

For this reason, in recent years, the definition of data quality has been at the centre of scientific debate, and several international organisations have participated actively to examining this question, contributing to a new definition of data quality and expanding the one commonly used in the past when quality was essentially identified with the concept of data accuracy.

The definition of quality that is now internationally recognised can be seen as a more multifaceted concept covering a wider spectrum of issues. It can be defined as 'fitness for use' in terms of user needs, i.e. that the most important quality characteristics depend on user perspectives, needs and priorities, which vary across groups of users (Organisation for Economic Co-operation and Development [OECD], 2011). This means that even if the data are accurate, they may not be of good quality if produced too late, if they are not easily accessible or if they appear to be in conflict with other data. In other words, to evaluate the quality of the data, it is necessary to evaluate several dimensions related both to the output/product and to the statistical process itself.

According to Eurostat (Eurostat, 2014a), output/product quality can be assessed by exploring five different dimensions: relevance, accuracy and reliability, timeliness and punctuality, coherence and comparability and accessibility and clarity. Relevance measures the degree to which statistical information meets current and potential user needs (Eurostat, 2014a). For this reason, measuring relevance is subjective and depends upon varying user needs. Accuracy refers to the degree to which the data correctly estimate or describe the phenomenon they are designed to measure, whereas reliability refers to the confidence that users have in a product based simply on their image of the data producer (OECD, 2011). Moving to the third dimension, timeliness can be defined as the length of time between data availability and the event or phenomenon the data describe, whereas punctuality refers to the time lag between the actual delivery of data and the target date on which they were scheduled for release as announced in an official release calendar (Eurostat, 2014a). The coherence of data products is the degree to which they are logically connected and mutually consistent; comparability measures the impact of differences in applied statistical concepts, measurement tools and procedures where statistics are compared between geographical areas or over time. Finally, accessibility and clarity reflect the simplicity and ease with which users can access, use and interpret statistics with the appropriate supporting information and assistance. Looking at the process itself, it is important to remember that output quality can be achieved through a high-quality process via effective and efficient management of statistical processes (IMF, 2003).

[1] University of Brescia, Italy.

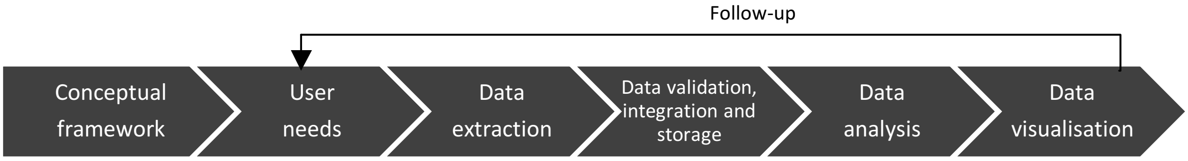
Against this background, this paper begins with a brief presentation on the Systemic Risk Tomography: Signals, Measurements, Transmission Channels and Policy Interventions (SYRTO)[2] project, and in particular, describes the task of defining and implementing the data centre. Then, it describes the statistical processes and organisational measures adopted in order to meet the quality dimensions listed above within the working group for the collection and integration of data from different sources into a single database that researchers involved in the project can use.

1. The SYRTO project: providing data to prevent, manage and resolve systemic crises in the Eurozone

The SYRTO project is funded by the European Union (EU) under the 7th Framework Programme; it aims to create an early warning system to identify potential threats to financial stability and realise an ensemble of suggestions and prescriptions on the appropriate policy measures, governance structure and macro-prudential supervision to prevent, manage and resolve systemic crises in the Eurozone. Suggestions and prescriptions cannot be formulated without a solid and robust data analysis, so one of the main tasks of the SYRTO project is defining and implementing a data centre to collect all of the relevant information to monitor markets, financial institutions and the economy and to evaluate the severity of the risks impact (both individual and systemic risks).

The project aims to define a statistical process that includes the steps of data collection, integration, validation and analysis, with the ultimate goal of releasing statistics so that policy makers, researchers and society in general interested in the project's subject matter can use them. Figure 1 summarises this statistical process, from the conceptual framework to the final data visualisation.

Figure 1. SYRTO data process orientation

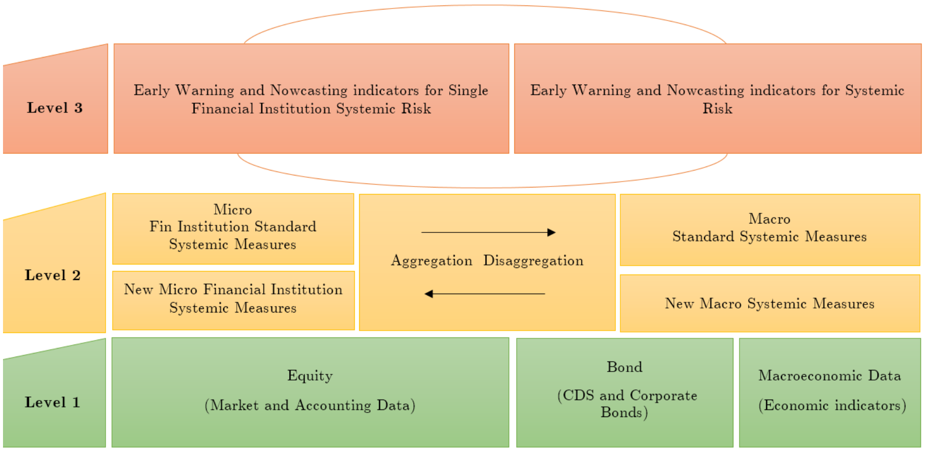


Regarding the conceptual framework (figure 2), the researchers involved in the project created the SYRTO data framework to bring modularity and flexibility to the database with low time-consuming tasks, which allows for modification at a single level (Costola, 2014). In detail, the SYRTO data framework is composed of three hierarchical levels:

- Level 1 includes "raw" financial and economic data.
- Level 2 includes systemic risk measures estimated on the basis of the first-level data.
- Level 3 includes early warning and nowcasting indicators.

[2] This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n° 320270. www.syrtoproject.eu

Figure 2. SYRTO data framework


Level 3
Early Warning and Nowcasting indicators for Single Financial Institution Systemic Risk
Early Warning and Nowcasting indicators for Systemic Risk
Level 2
Micro Fin Institution Standard Systemic Measures
New Micro Financial Institution Systemic Measures
Aggregation Disaggregation
Macro Standard Systemic Measures
New Macro Systemic Measures
Level 1
Equity (Market and Accounting Data)
Bond (CDS and Corporate Bonds)
Macroeconomic Data (Economic indicators)

Source: Costola (2014)

From analysing the theoretical framework from the point of view of the data process shown in Figure 1, “raw” data go through the early stages of the process (data extraction, integration and validation). The researchers develop systemic risk measures, early warning and nowcasting indicators during the data analysis phases.

That said, the following paragraphs focus on the first steps of the SYRTO data process to explore the statistical processes and organisational measures adopted in order to meet the quality dimensions listed in the first paragraph for the definition of user needs, data extraction, integration and validation.

2. Definition of user needs and evaluation of data sources

The starting point for defining the SYRTO data process is to identify the user needs and then map the existing data sources that can be used to respond to these needs. This is also done to determine whether it is necessary to implement new data collection processes, e.g. surveys, to collect data that are not available in existing data sources.

As already mentioned, the aim of the data centre is to collect data that researchers from the SYRTO project can use to estimate measures of systemic risk and develop early warning and nowcasting indicators. For this reason, at this stage of mapping, not only the group of statisticians directly involved in developing the data centre but the whole group of researchers working on the project are involved; this is necessary to help identify the needs of the whole group not only in terms of data content but also in terms of accuracy, coverage, frequency, timeliness, metadata required for interpretation and the relationship to other relevant statistical outputs (Eurostat, 2014a).

That said, keeping in mind the aims of the SYRTO project, data requirements refer to data about sovereigns, banks, other financial intermediaries and corporate and market data (equity, bond, derivatives, interest rates). In detail, they include data on single financial institutions and relevant macroeconomic variables for contagion and systemic risk analysis, which can be classified into three categories (Costola, 2014):

- Macroeconomic data (economic indicators and sensitive variables).
- Bonds (sovereign and corporate bonds and credit default swaps [CDSs]).

- Equity (market and accounting data).

On the sovereign side, macroeconomic data and bond data covers the European aggregates (if relevant), the 28 EU member countries[3] and, if data are available, other relevant markets such as Japan, Norway, Switzerland and the US starting from 1 January 1996. On the corporate side, bond and equity data focus on the constituents of the STOXX Europe 600 Index starting from 1 January 1990. Looking at the possible data sources, these three categories can be extracted from existing data sources, and in some cases, more than one data source is available. Possible data sources taken into account can be divided into three groups:

- Data released by international organisations.
- Data released by national agencies.
- Data belonging to databases developed by private companies.

In assessing the possible data sources, the research group has established a priority order of sources. Primary sources for data acquisition are data that are released by international organisations. Considering that data released by international organisations have mostly been collected from national agencies in the first place, data can potentially be taken from the international organisation or from the national agency. Priority is given to data from international organisations because they have the advantage of benefiting from any editing or compilation already undertaken by the international agency collecting the data, e.g. harmonisation of the data, and the costs of data extraction are reduced compared to extracting data from national sources (OECD, 2011). If the same data are released by more than one international organization, priority is given to data from Eurostat[4] and the European Central Bank (ECB) since they have an official mandate to provide statistics at European level that enable comparisons between countries and regions. If data are not released by Eurostat and ECB, they are taken from other international organisations (such as OECD, the World bank and IMF). In the case that data are not issued by international organisations, they are taken from the national agencies. Extracting data from databases developed by private companies is considered solely in the event that the data are not issued to the public from the first two sources.

In identifying data sources, the coherence, timeliness, accuracy, accessibility and interpretability of the data are evaluated. As regards the interpretability of data, particular attention is paid to the structure and the degree of completeness of metadata. Given the international nature of the project, existing international statistical guidelines and recommendations are used for concepts, definitions, units, classifications, nomenclatures and compilation methods. Divergences from these international standards are documented and justified. On-going statistical activities are reviewed at regular intervals in close partnership with stakeholders to identify new needs, to adopt the most appropriate statistical methods and to use the most effective technical solutions.

[3] https://www.ecb.europa.eu/stats/money/long/html/index.en.html

[4] www.eurostat.eu

2.1. Macroeconomic data

Going into detail, the first category of data refers to macroeconomic data, which can be classified into economic indicators and sensitive variables for banking systems and the financial sector.

Taking into account user needs and the project's aim, the economic indicators include monthly and quarterly statistics (table 1) measuring economic and employment developments starting from 1 January 1996. Looking at the data sources, data for European aggregates and EU member countries are all officially released by Eurostat in its online database[5]. For some of the variables of interest, Eurostat also releases data for non EU member countries (Japan, Norway, Switzerland and US). In case data for non EU member countries are not available in the Eurostat database (or in case they are not complete), data are taken from OECD online database[6]. The selected economic indicators cover seven different areas:

- *National accounts*. A coherent and consistent set of macroeconomic indicators, providing an overall picture of the economic situation covering gross domestic product (GDP) and its main components. For all the countries, excluding Japan, data are compiled according to the 2010 European system of national and regional accounts (ESA 2010). For US coverage is usually limited to key aggregates.
- *DEBT OVER GDP*. The total gross debt at nominal value outstanding at the end of each quarter for the general government sector.
- *Inflation*. Economic indicators that measure the change over time of the prices of consumer goods and services acquired by households. For all the countries, excluding Japan, data refer to Harmonised Indices of Consumer Prices (HICPs) released by Eurostat. HICPs give comparable measures of inflation for the countries and country groups for which they are produced.
- *House prices*. House Price Indices (HPIs) measuring inflation in the residential property market. At the moment, data not available for Japan, Norway, Switzerland and US.
- *Short-term business statistics* (STS). Short-term statistics give information on a wide range of economic activities according to NACE Rev.2 classification. Data refer to production, turnover and Producer Price Indices (PPI).
- *Labour force survey* (LFS). Household sample survey, providing data on labour participation of people aged 15 and over and on persons outside the labour force.
- *Business and consumer surveys* (BCSs). Data on current economic activity and its perspectives based on the opinion of economic actors, such as entrepreneurs and consumers. Harmonised data are available for EU aggregates and EU member countries. Data are not available for Norway.

[5] http://ec.europa.eu/eurostat/data/database

[6] http://stats.oecd.org/

Table 1. List of the economic indicators to be included in the SYRTO database

Area	Variables	Frequency	Eurostat	OECD
National accounts	Gross domestic product at market prices Final consumption expenditure Household and NPISH final consumption expenditure Final consumption expenditure of general government Gross capital formation Gross fixed capital formation Exports of goods and services Imports of goods and services Gross value added (at basic prices) Real labour productivity per person employed Nominal unit labour cost	quarterly	European Union, Euro area, EU Member States, Norway, Switzerland, US	Japan
Debt/Gdp	Debt/Gdp	quarterly	European Union, Euro area, EU Member States	-
Inflation	All-items Harmonised Index Price Consumer (HICP; 2015=100) Overall harmonised index excluding energy and unprocessed food (2015=100)	monthly	European Union, euro area, EU Member States, Norway, Switzerland, US	-
	CPI: all items (Index 2010=100) CPI: all items non-food non-energy (Index 2010=100)	monthly	-	Japan, Switzerland
House prices	House price index	quarterly	European Union, euro area, EU Member States, Norway	-
Short-term business statistics - Industry	Production Turnover Producer Price Index (PPI) Turnover	monthly	European Union, euro area, EU Member countries, Norway	Japan, Switzerland and US
Labour Force Survey	Total employment (15 years and more) Activity rate (15 to 64 years) Activity rate (15 to 24 years) Total unemployment rate (15 to 74 years) Youth unemployment rate (15 to 24 years)	quarterly monthly	European Union, Euro area, EU Member countries, Norway, Switzerland	Japan, US
Business and consumer survey	Production development over the past 3 months Production expectations over the next 3 months Economic sentiment indicator Industrial confidence indicator General economic situation over the last 12 months General economic situation over the next 12 months Consumer confidence indicator	monthly (Switzerland quarterly)	European Union, Euro area, EU Member countries	Japan, Switzerland, US[7]

Moving to the sensitive variables (Table 2), this category includes indicators for financial and banking systems covering the Euro area and such as the Composite Indicator of Systemic Stress (CISS) used by the European Central Bank (ECB) and its constituents (Hollò et al., 2012), and data on daily liquidity conditions[8].

[7] Data for Japan, Switzerland and US are taken from OECD database and do not cover Economic sentiment indicator, General economic situation over the last 12 months and, limited to Japan, General economic situation over the next 12 months.

[8] The data on banks' access to the ECB marginal lending facility can be downloaded here: Data on daily liquidity conditions, www.ecb.europa.eu/stats/monetary/res/html/index.en.html

Table 2. List of the sensitive variables to be included in the SYRTO database

Variables	Frequency	Coverage	Source
Composite Indicator of Systemic Stress (CISS)	daily	Euro area	ECB
CISS constituents	daily	Euro area	ECB
Sovereign Systemic Stress Composite Indicator	monthly	Austria, Belgium, Czech Republic, Denmark, Finland, France, Germany, Greece (GR), Hungary, Ireland, Italy, Netherlands, Poland, Portugal, Spain, Sweden, United Kingdom (GB)	ECB
Daily liquidity conditions	daily	Euro area	ECB

2.2. Bond data

Moving to the bond data, this second category includes data about sovereign and corporate bonds and CDSs.

On the sovereign side, data refer to short-term and long-term interest rates covering EU aggregates, the EU member countries[9], Japan, Norway, Switzerland, the US and the CDSs for all of the aforementioned countries. For short-term interest rates, data refer to 3-month interest rates on the financial market for loans or deposits. Data after 1999 cover euro area, EU27 and national series for the EU Member States which are not members of the euro area. Before 1999, national series for all EU Member States are available. Looking at the long-term interest rates for the European aggregates and EU member countries, the data refer to Maastricht criterion bond yields, which are long-term interest rates used as a convergence criterion for the European Monetary Union (EMU) and based on the Maastricht Treaty. For Japan, Norway, Switzerland and the US, data refer to government bonds with outstanding maturities of 10 years.

Regarding sovereign CDSs, the CDS market is an over-the-counter (OTC) market almost entirely populated by institutional investors; therefore, in contrast with an organised exchange, there are no formally established clearing and settlement mechanisms providing reliable information on prices. The information on prices must be gathered from market participants on the basis of their voluntary participation in periodic surveys, with all the potential shortcomings such a situation may incur (Mayordomo et al., 2014). For this reason, it is easy to understand why international organisations or other official statistics providers do not officially release data on CDSs. Data on CDSs can be collected by accessing one of the several private databases specialising in economic and financial data.

Sovereign CDSs data collected for the SYRTO project refer to CDSs with maturities of three years, five years, seven years and ten years. The type of contract chosen is complete restructuring (CR). The full-restructuring clause was the standard contract term in the 1999 ISDA[10] credit derivatives definitions. Under this contract option, any restructuring event qualifies as a credit event (and any bond of maturity of up to 30 years is deliverable).

CDS data are obtained from the Thomson Reuters Datastream database. Datastream provides CDS data based on two sources: CMA[11] and Thomson Reuters[12]. Both sources collect information on prices, maturities and restructuring types from markets with 30 or more contributors such as asset managers,

[9] Excluding Estonia. For more information: http://ec.europa.eu/eurostat/cache/metadata/en/irt_lt_mcby_esms.htm

[10] The International Swaps and Derivatives Association. www.isda.org

[11] CMADatavision. www.cmavision.com

[12] http://thomsonreuters.com/en.html

banks or hedge funds. From CMA, CDS spreads beginning in 2004 can be obtained, but the series are no longer accessible after October 2010. The second source, Thomson Reuters, reports CDS data up to the present, but CDS series for most countries are only available from the end of 2008 onward. So, in order to obtain long-term series, data from the two sources were appended. The Thomson Reuters CDSs are denominated in US dollars, whereas CMA CDSs are released for some countries in US dollars and for others in Euros, but the change in currency for a single series is not problematic since CDS data are measured in basis points and are therefore free of units.

On the corporate side, data refer to five years CDSs on a sample of European supervised banks by the European Central Bank[13] for which data were available in the Datastream database.

Table 3. List of bond data to be included in the SYRTO database

Variables	Frequency	Coverage	Source
Long-terms interest rates - Maastricht criterion bond yields	monthly	Euro area and EU member countries	ECB
Long-terms interest rates - 10 yrs government bonds	monthly	Japan, Norway, Switzerland, USA	OECD
Short-terms interest rates - 3-month money market interest rates	monthly	Euro area, EU members countries not included in the Euro area, Japan, USA	Eurostat
Short-terms interest rates Money market interest rates - 3 months	monthly	Norway, Switzerland	OECD
Sovereign CDSs (3YRS, 5YRS, 7YRS, 10YRS maturity)	daily	EU member countries Japan, Norway, Switzerland, USA	Thomson Reuters Datastream
Monetary aggregates – M3	daily	EU member countries excluding Portugal and Croatia	Thomson Reuters Datastream
Corporate CDS – European Sistematically important banks	daily	30 banks monitored by ECB	Thomson Reuters Datastream

2.3. Equity data

Finally, looking at the last category of data, equity data include both daily equity data on the constituents of the STOXX Europe 600 Index, accounting data of non-financial companies from countries covered by Eurostoxx 600 and a selected list of daily Datastream Global Equity Indices.

Starting from the STOXX Europe 600 Index components, daily equity data are taken from the Thomson Reuters Datatstream starting from 1 January 1990.
Annual accounting data are extracted from the Bureau Van-Dijk Amadeus database[14], a database containing accounting data on over 14 million companies across Europe from 2007 onward.
Also, a list of selected Datastream Global Equity Indices covering European markets broken down into different sectors has been defined. The Global Equity Indices cover EU member countries, excluding Estonia, Lithuania, Latvia, Slovakia and Romania, since Datastream does not cover them, and also other relevant markets such as Japan, Norway, Switzerland and the US.

Datastream Global Equity Indices provide a range of equity indices that are calculated on a representative list of stocks for each market. The number of stocks for each market is determined by the size of the market. The sample covers a minimum of 75–80% of total market capitalisation. Within

[13] https://www.bankingsupervision.europa.eu/banking/list/who/html/index.en.html
[14] www.bvdinfo.com/en-gb/our-products/company-information/international-products/amadeus

each market, stocks are allocated to industrial sectors using the Industry Classification Benchmark (ICB) jointly created by the Financial Times and the London Stock Exchange (FTSE) and Dow Jones. The sector indices are then calculated and index constituents for each market are reviewed quarterly.

Table 4. List of equity data to be included in the SYRTO database

Variables	Frequency	Coverage	Source
Equity data	daily	STOXX Europe 600 components	Thomson-Reuters Datastream
Accounting data	annual	Companies from countries included in the STOXX Europe 600	Bureau Van-Dijk Amadeus
Datastream Global Equity Indices	daily	EU Member countries, (excluding Estonia, Lithania, Latvia, Slovakia and Romania), Japan, Norway, Switzerland, USA	Thomson-Reuters Datastream

3. The SYRTO data center: data extraction, integration and storage procedures

After having identified the user needs, next step is the definition and implementation of a system to manage the data process. Looking at the data quality, it has to be kept in mind that in this process, the steps have been evaluated with respect to the objectives for which the statistical activity was initiated and also that in a modular process such as the one shown in Figure 1, the quality of the data depends on the quality of the data collected but also on the quality of the internal processes for collection, processing, analysing and disseminating the data and metadata.

From the operational point of view, the data process is developed using the Konstanz Information Miner Platform (KNIME) (Berthold et al., 2008; www.knime.org). KNIME is an open-source platform based on the Eclipse Platform, which allows the user to easily and intuitively manage modular data analysis environments. It also facilitates quick and interactive changes to the analysis process and enables the user to visually explore the results. Thanks to additional plugins, it is also possible to run R scripts and Matlab scripts in KNIME, giving the user access to a wide library of statistical routines and models.

3.1. Data and metadata extraction

As noted in the previous section, there are three main data sources: international agencies, the Thomson Reuters Datastream database and the Bureau Van-Dijk Amadeus database. While working with these sources, it is important to remember that procedures for data and metadata extraction are an important phase of the data process that must be defined and implemented in order to ensure a high level of quality in terms of accuracy, credibility, timeliness, accessibility and interpretability without forgetting efficacy and economic efficiency. For this reason, data extraction procedures should be automated to avoid manual imputation and intervention. Data extraction procedures must also be error-proofed. Possible sources and types of error are analysed, and provisions are put in place to check and correct for errors. In case procedures cannot be automated, a research group should data share and validate them. Also, according to the OECD (2011), when defining the procedures for data extraction, it is better to gather all the information necessary to assess the aspects related to data quality, including:

- Presence of provisional or non-verified information on available data and metadata
- Expected future revisions in data and metadata
- Potential problems in data quality for data not completely documented by related metadata
- Any other information useful to evaluate aspects related to data quality

Looking at the SYRTO project, due to the characteristics or limits of the data sources, the degree of automation of the procedures for data extraction varies from source to source.

Starting from international agencies' databases the first category of data sources, the extraction procedures ensure high-quality standards in terms of output; this is mainly due to the high quality of the data guaranteed by the data providers and the efficiency and effectiveness of the procedures for data and metadata extraction.

Eurostat, the OECD and the ECB officially release the data of interest, making them available to the public in the three institutions' online databases. Data can be extracted using the Statistical Data and Metadata Exchange (SDMX) standard[15], a statistical and technical standard to efficiently harmonise and disseminate statistical data. The SDMX is an initiative that seven international organisations began promoting in 2001: the Bank for International Settlements (BIS), the ECB, Eurostat, the International Monetary Fund (IMF), the OECD, the United Nations (UN) and the World Bank (WB). They aimed to develop and implement more efficient processes for exchanging and sharing statistical data and metadata among international organisations and their member countries.

The aim of the SDMX initiative is to create and maintain technical and statistical standards and guidelines, together with an information technology architecture and information technology tools, to be used and implemented by the SDMX sponsors and other organisations that use statistical data and exchange metadata (Salou and Sosnovsky, 2010). Combined with modern information technology, these SDMX standards and guidelines should improve efficiency when managing statistical business processes. In other words, the aim of the SDMX initiative is to establish a set of commonly recognised standards, making it possible not only to have easy access to statistical data but also to metadata that makes the data more meaningful and usable (UN Statistical Commission, 2015). In this way, several quality dimensions such as timeliness, accessibility, interpretability, coherence and cost efficiency can be improved through the use of SDMX standards.

[15] The list of SDMX cross-domain concepts and related code lists are available from the SDMX website: http://www.sdmx.org

That said, extracting the data of interest of the research group was achieved by using SDMX queries through SDMX web services, which grant access to Eurostat's, the ECB's and the OECD's online data warehouses. Specifically, the SDMX queries have been implemented through the use of RSDMX, an R package to parse and read SDMX documents in R that provides a set of classes and methods to read data and metadata documents exchanged through the SDMX framework (Blondel, 2014). Being developed in R, routines for data extraction can be coded directly into the data management process developed in the KNIME platform, ensuring high standards in terms of process efficiency and effectiveness.

Moving to the Thomson Reuters Datastream database, the extraction procedures ensure quite good-quality output standards in terms of the efficiency and effectiveness. Automated procedures to extract data from such a data source have been defined using the Advance for Office (AFO) add-ins[16], a user interface integrated in Microsoft Excel provided within Datastream; it allows the user to define automated TSI requests that can be repeated over time without the need to manually edit the details of such queries. Compared to the procedures described for the data extracted from international agencies' databases, procedures for extracting data from Datastream have the weak point of not being able to be directly included in one node of the data process defined in KNIME. This necessitates two steps: firstly extraction of data using AFO add-ins and secondly importing the Excel series into KNIME.

Finally, the procedures for extracting data from the Bureau Van Dijk-Amadeus database currently provide lower-quality standards in terms of efficiency and effectiveness. These are mainly lacking because of the lower degree of automation of procedures for extracting data from this database. For characteristics of the database, especially related to the limits imposed during the download of the data, it is possible to define automatic procedures for the selection of the variables of interest; however, the user cannot define the procedures for automatic download data. Currently, the Amadeus database is the only source available to the research group in terms of accessing financial information for European companies. The research group is still considering the possibility of developing more automated procedures.

3.2. Data integration and storage

The process of data integration and storage is defined and managed in KNIME. The aim is to manage all of the data through time-series objects, i.e. data that have the property of being temporally ordered, each of which contains one of the variables discussed in the previous sections not only in terms of data but also metadata, which may help the user to more easily and more quickly understand the characteristics of the series.

From the practical point of view, thanks to the integration of R in KNIME, lists of xts objects (Ryan and Ulrich, 2012) have been defined, each of which contains the data and metadata related to one of the variables described in the previous paragraphs. Among the many classes of time-series objects available in R, the choice to define xts objects lies in its simplicity of use, its overall flexibility and in the possibility to define user-added attributes. An xts object can be considered an extension of a zoo object (Zeileis and Grothendieck, 2005), differing from the zoo class in three key ways: the use of formal time-based classes for indexing, internal xts properties and user-added attributes. Simplified, an xts object contains an array of values comprising the data (often in matrix form) and an index attribute to provide information about the data's ordering (Ryan and Ulrich, 2012).

[16] For more information: http://extranet.datastream.com/User%20Support/PubDoc/Advance.htm

The possibility of defining user-added attributes is a feature the importance of which should not be underestimated. With this feature, it is possible to maintain a single object within the values of the series in its metadata, providing the user with detailed information that can simplify data use. Indeed, for each single xts object, the basic metadata are specified (such as the complete name of the variable, the data source, the frequency, etc.) and a link to the complete metadata, if available, is given.

The process of defining, gathering and storing xts objects is handled automatically and with practically zero manual intervention. Automatic procedures are adopted to constantly monitor how data storage is progressing, how the coverage of the data received compares to expectations and how significant any data revisions are in order to take timely action to solve any emerging problems.

4. A quality assessment from the SYRTO data processing view

In the previous paragraphs, the issue of quality of both product and process underlied the definition and the implementation of all the steps of data extraction, integration and storage. The quality strategy has been implemented throughout the entire SYRTO data process to identify the aspects of quality relevant for the SYRTO data centre and to find proper measures for their assessment. As a consequence of this effort, quality indicators are considered at each stage of the process of data extraction and integration (Table 5). They are monitored to control data quality and to provide information to improve the process. Some are traditional indicators (Eurostat, 2014b), useful mainly as documentation to assess the quality of the process and output. Others are more oriented to carefully monitor every step of the process.

Table 5. Quality dimensions, indicators and measures for the different phases of the SYRTO data process

Step	Quality dimension	Indicator	Measure/Method
Definitions of user needs	Relevance	Data completeness	Gap between key user interests and released statistics in terms of concepts, coverage and detail
Data extraction	Accessibility and clarity	Data accessibility	Data are made available to all the users and they are presented in a way that facilitates interpretation and meaningful comparisons
		Metadata accessibility	Documentation on concepts, scope, classifications, basis of recording, data sources, and statistical techniques is available, and differences from internationally accepted standards, guidelines, or good practices are annotated
Data integration and storage	Accuracy	Missing data rate	the rate for a given variable, which is defined as the ratio between the number of periods for which the data are not available and the total number of periods observed
	Timeliness and punctuality	Time lag - data integration	The length of time (number of days) between the reference period and the date of data integration into SYRTO database

Conclusion

This paper provides an overview of the user needs and the process of data extraction and integration defined within the SYRTO project with particular regard to the procedures adopted to ensure the highest possible level of quality in terms of both product and process. This is done via adopting special tools and methods for data extraction and integration, which are identified in terms of indicators currently implemented to monitor the evolution of the project and ensure that the process is developed with respect to the objectives for which the statistical activity was initiated. The procedures presented in this paper are still being implemented, and some of them will be finalised only in the coming months; therefore, the present paper should be considered a work in progress, the final version of which will be released only when the data centre is successfully implemented.

References

Berthold, M.R. et al. (2008). KNIME: The Konstanz Information Miner. In Preisach, C. et al. (eds.) Data analysis, machine learning and applications: studies in classification, data analysis, and knowledge organization, Vol. V, pp. 319–326.

Blondel, E. (2014). Rsdmx – Tools for reading SDMX data and metadata documents in R – version 0.4-2. Zenodo. 10.5281/zenodo.11551

Costola, M. (2014). SYRTO Data Management. Presentation made at the *Workshop on Systemic Risk Policy Issues for SYRTO*. Head Office of Deustche Bundesbank, Guest House. Frankfurt am Main. July, 2 2014

Eurostat (2009). ESS standard for quality reports. Available online at: http://epp.eurostat.ec.europa.eu/portal/page/portal/ver-1/quality/documents/ESQR_FINAL.pdf

Eurostat (2014a). ESS handbook for quality reports 2014. Available online at: http://ec.europa.eu/eurostat/documents/64157/4373903/01-ESS-Handbook-for-Quality-Reports-2014.pdf/d6152567-a007-4949-a169-251e0ac7c655

Eurostat (2014b). ESS quality and performance indicators 2014. Available online at: http://ec.europa.eu/eurostat/documents/64157/4373903/02-ESS-Quality-and-performance-Indicators-2014.pdf/5c996003-b770-4a7c-9c2f-bf733e6b1f31

Hollò, D., Kremer, M., and Lo Duca, M. (2012). CISS – A composite indicator for systemic stress in the financial system. ECB – Working Paper Series, 1426.

IMF (2003). Data quality assessment framework and data quality program. Available online at: www.imf.org/external/np/sta/dsbb/2003/eng/dqaf.htm

Mayordomo, S., Peña, J.I., and Schwartz, E. (2014). Are all credit default swap databases equal? European Financial Management, Vol. 20, Issue 4, pp. 677–713. doi: 10.1111/j.1468-036X.2013.12023.x

OECD (2011). Quality framework and guidelines for OECD statistical activities. Version 2011/1. Available online at: www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=std/qfs(2011)1&doclanguage=en

Ryan, J.A., and Ulrich, J.M. (2012). Xts: eXtensible Time Series. R package version 0.8-6, URL http://CRAN.R-project.org/package=xts .

Salou, G., and Sosnovsky, X. (2010). SDMX as the logical foundation of the data and metadata model at the ECB: the IFC's contribution to the 57th ISI Session, Durban, South Africa, August 2009. IFC Bulletin, No. 33. Available online at: www.bis.org/ifc/publ/ifcb33.htm

UN Statistics Division (2003). Handbook of statistical organization, third edition: the operation and organization of a statistical agency, Series F, No. 88.

UN Statistical Commission (2015). Report of the statistical data and metadata exchange sponsors, E/CN.3/2015/33. Available online at:
http://unstats.un.org/unsd/statcom/doc15/2015-33-SDMX-E.pdf

Zeileis, A., and Grothendieck, G. (2005). Zoo: S3 infrastructure for regular and irregular time series. Journal of Statistical Software, Vol. 14, Issue 6.