

End users' scenario for OS BI Statistical Data Life Cycle Management

Author:

Manuel Da Silva¹, Philippe Petit¹

¹*Artemis Information Management SA*

This work is licensed under the Creative Commons Attribution-Share Alike 3.0 License.

To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/3.0/> or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.

OW2 BI Initiative deliverable	Title: End users' scenario for OS BI: Statistical Data Life cycle Management	January 29 th , 2009
Author: Artemis	Revision: Engineering	Page: 1/12

Revision History

Version	Lead Author	Summary of Changes	Date
1.0	MDS, Artemis	First version release	Jan, 29 th , 2009

Table of Contents

Abstract3

About the BI Initiative4

1. Introduction.....5

2. Statistical data user needs6

 2.1 Timeliness6

 2.2 Data coherence and completeness.....6

 2.3 Data comparability and quality6

 2.4 Adjustment, completion and correction7

 2.5 Confidentiality treatment7

 2.6 Data analysis and dissemination7

 2.7 Management of metadata8

 2.8 Communication8

3. Statistical DLM scenario9

 3.1 Monitoring of the data reception.....9

 3.2 Detection of the file format9

 3.3 Checking of the content of each field9

 3.4 Production of quality reports.....10

 3.5 Treatment of revised data.....10

 3.6 Data dissemination11

Conclusions.....12

OW2 BI Initiative deliverable	Title: End users' scenario for OS BI: Statistical Data Life cycle Management	January 29 th , 2009
Author: Artemis	Revision: Engineering	Page: 2/12

Abstract

Statistics are more and more important for policy makers and economists. It is essential to have reliable data on which statistics are based.

Open Source BI platforms and tools can be very effective solutions in order to support statistical data analysis, mostly because they evolve much faster than commercial platforms since they are not constrained by compatibility problems and rigid (or even obsolete) architectures.

This document investigates the statistical Data Life cycle Management (DLM) as a prominent end users' scenario for OS BI.

The first part of the paper describes the needs of the different actors intervening in the statistical data life cycle. A second part is dedicated to the description of a statistical DLM scenario. In conclusions, a proposal for a next step is given.

OW2 BI Initiative deliverable	Title: End users' scenario for OS BI: Statistical Data Life cycle Management	January 29 th , 2009
Author: Artemis	Revision: Engineering	Page: 3/12

About the BI Initiative

OW2 Initiatives are joint efforts of OW2 Members and not-OW2 Members aiming at facilitating the use of OW2 technologies by mainstream systems integrators, end-users, academia and software vendors. Within an Initiative, participants work together to develop both technical integration between projects and business synergies in order to address specific market needs.

The BI Initiative is a join effort set up to:

- improve the coordination effort in the OS BI context
- increase the use of OS BI solutions at enterprise level
- strengthen connections between integrators, vendors, users and the research communities
- attract more attention from the research activities to foster innovative BI solutions and practices.

Main activities are:

- integration of a full stack of OS BI solutions and tools
- promotion of a service network to support the entire stack
- kick-starting a community contributing to strengthen the current OS BI platforms
- creation of a research network on BI topics, fostering a closer cooperation between researchers and OS BI vendors, to promote the development of a new generation of OS BI platforms.

Participants of the BI Initiative are:

- Altic, France (www.altic.org)
- Artemis Information Management, Luxembourg (www.artemis.lu)
- ClaraVista, France (www.claravista.fr)
- Engineering Ingegneria Informatica, Italy (www.eng.it)
- Ingres, USA (www.ingres.com)
- Talend, France (www.talend.com);
- DEIS-University of Bologna-Italy (www.eng.unibo.it/PortaleEn/default.htm)
- DTI-University of Milan (<http://ra.crema.unimi.it>), Italy
- DISI-University of Trento (www.dit.unitn.it), Italy
- OW2 individual members.

More information about OW2 BI Initiative: www.ow2.org/view/BusinessIntelligence/

More information about OW2 Consortium: www.ow2.org

OW2 BI Initiative deliverable	Title: End users' scenario for OS BI: Statistical Data Life cycle Management	January 29 th , 2009
Author: Artemis	Revision: Engineering	Page: 4/12

1. Introduction

Statistics are more and more important for policy makers and economists. It is essential to have reliable data on which statistics are based. In order to achieve the best level of reliability, it is really important to put very strong efforts on Data Life cycle Management (DLM), i.e. on managing the flow of an information system's data throughout its life cycle: from the collection to the dissemination of the data.

Everything starts with the implementation of the data collection. The statisticians decide what they want to collect and define all concepts allowing data providers to understand what they should transmit. Many factors linked to the data collection are defined: variables to be provided, periodicity, timeliness, transmission format, transmission mean, etc. Once data collection implemented and started, data have to be checked and validated using statistical concepts more or less complex. After validation, data need to be analysed in more details in order to define the appropriate indicators to be disseminated. Finally, data are disseminated in various formats: electronic versions (PDF documents, Web portal) or printed versions (e.g. leaflets, pocketbooks)

The first part of this document will describe the needs of the different actors intervening in the statistical data life cycle. A second part will be dedicated to the description of a statistical DLM scenario.

OW2 BI Initiative deliverable	Title: End users' scenario for OS BI: Statistical Data Life cycle Management	January 29 th , 2009
Author: Artemis	Revision: Engineering	Page: 5/12



2. Statistical data user needs

At each different step of statistical data life cycle, data users express specific needs.

2.1 Timeliness

Timeliness refers to data availability, whether for dissemination or for further processing. It is measured with respect to the time lag between the end of the reference period and the release of final data or of the first provisional results.

☞ **The user needs to assess timeliness of data transmission in order to identify the data providers not meeting the deadlines and to communicate them reminders.**

Timeliness is a crucial element of data quality: adequate timeliness corresponds to a situation where policy-makers can take decisions in time to achieve the targeted results. Usually, data are not released immediately at the end of the period they refer to, since data collection, data processing and data dissemination work needs to be performed. It sounds essential to reduce the time between the end of the reference period and the dissemination of the data.

2.2 Data coherence and completeness

When originating from different sources, and in particular from statistical surveys using different methodology, statistics are often not completely identical, but show differences in results due to different approaches, classifications and methodological standards. In order to have coherent statistics, the statisticians implement methodologies to be followed by all data providers. These methodologies mainly consist in defining the different variables to be collected and establishing the codifications to be used.

☞ **It is essential to be able to verify that all data providers comply with the rules established in the methodologies. If not, they have to be informed of the problems encountered.**

For instance, all variables required have to be provided guaranteeing the completeness of the information. It is also important to check that the correct codifications have been used in order to be able to interpret correctly the transmitted data. The verification of the data coherency is a first important step in data validation. Indeed, without coherency, data are not comparable.

2.3 Data comparability and quality

Although the data coherency is achieved, this does not mean that data are comparable because of different interpretations of the methodologies. A second important step of data validation consists in measuring the impact of methodological differences on the comparison of statistics between different data providers, over time or across statistical domains.

☞ **The statistician needs to implement his statistical methods allowing to assess the quality of data and to visualize the results in reports highlighting the outliers. These reports are produce on regular basis and are circulated to the different actors with accompanied comments.**

OW2 BI Initiative deliverable	Title: End users' scenario for OS BI: Statistical Data Life cycle Management	January 29 th , 2009
Author: Artemis	Revision: Engineering	Page: 6/12

Usually, the methods used to assess the data quality rely on simple mathematical formulas and do not require the usage of specific statistical tools.

2.4 Adjustment, completion and correction

Visualisation of the results of the data quality analysis is important but is not sufficient. Indeed, once the bad quality issues are identified, they have to be corrected. These corrections can be applied through different ways: data providers send revised data replacing the previous ones or the statistician apply corrections to the data. The reception of revised data or corrections by the statistician can imply three situations:

- a) Values are modified compared to the previous sending (adjustments or corrections, e.g. seasonal adjustments).
- b) New records have to be taken into account (completion).
- c) Records remain identical to the previous sending.



The user needs to keep track of the modifications applied. He should have access to the latest data in priority but should also be able to access all different versions of the data before and after corrections.

For instance, reports presenting the last data and indicating the modifications applied compared to the previous version of the data is very useful.

2.5 Confidentiality treatment

Confidentiality refers to a property of data with respect to whether, for example, they are public or their dissemination is subject to restrictions. For instance, data allowing identifying a physical or legal person either directly or indirectly are characterised as confidential. Often, there are procedures in place to prevent dissemination of restricted or confidential data, including aggregation rules when disseminating data. In fact, data can be confidential at different levels:

- Primary confidentiality: the single record is considered as confidential;
- Secondary confidentiality: possibility of deducing primary confidential data;



A sensitive issue is to establish automated procedures allowing treating the different levels of confidentiality in order to not disseminate data considered as confidential.

All data can be used for internal analysis purposes. The confidentiality concerns essentially the data dissemination.

2.6 Data analysis and dissemination

Once data validated, some simple or complex statistical calculations are applied on data (aggregation, growth rates or ANOVA, ACP, etc). The user usually creates standard reports presenting the data and the results of the calculations. These reports are produced regularly for the period of observation (e.g. month, quarter or year). The first need is to visualise the basic data at different levels of aggregation. Moreover, these data have to be usable to make either simple or complex calculations.

OW2 BI Initiative deliverable	Title: End users' scenario for OS BI: Statistical Data Life cycle Management	January 29 th , 2009
Author: Artemis	Revision: Engineering	Page: 7/12



☞ **The data and the results of the calculations should be presented in readable reports that are produced on regular basis, i.e. the reports are almost always identical depending on the reference period observed. The possibility to produce ad-hoc reports is also important for punctual analysis on very precise data.**

Another important issue concerns the comments accompanying the data reports. Indeed, the data can be considered as crude information and the explanations of the data turn this information into facts.

☞ **The results of the analysis should be presented in reports containing both data and comments. These reports are produce on regular basis and are disseminated.**

The circulation of the reports depends on the statistical analysis performed. Indeed, specific results may be circulated only to specific people. For example, basic data or simple calculations might be done available on the web or on paper publications whereas more complex will be used more restrictively.

☞ **For large public publications, a “pixel-perfect” quality is required.**

Apart from analysing data, an important role of the statistician is also to manage the information linked to data.

2.7 Management of metadata

Metadata are present in all statistical data life cycle. In a first step, metadata need to be completed or updated, mainly during the phase of data collection and validation. Then, metadata have mainly a consultative role during the analysis and the dissemination of data.

☞ **The metadata should be accessible during all steps of the statistical data life cycle either for consultation or modification.**

The more information on the data is available, the more enriching are the data. However, a large amount of metadata may become useless because simple information can be hidden in a huge volume of information and difficult to find.

☞ **Metadata should be "physically" linked to the data. For instance, when analysing some specific data, the specific metadata linked to these data should be extracted from the entire set of metadata.**

2.8 Communication

During all the statistical data life cycle, an important aspect is communication. Indeed, there are always several actors involved and everyone has to be informed with the right information and only the information he is concerned with. Typically, there are several exchanges containing comments on the reports produced or the results obtained.

☞ **It is essential to keep track of all the comments made, all the decisions taken during this multilateral communication and during all the steps of the data life cycle.**

OW2 BI Initiative deliverable	Title: End users' scenario for OS BI: Statistical Data Life cycle Management	January 29 th , 2009
Author: Artemis	Revision: Engineering	Page: 8/12

3. Statistical DLM scenario

The scenario presented below describes the data treatment of statistical data from the data reception to the data dissemination, passing through the analysis of the data.

3.1 Monitoring of the data reception


The data providers have to transmit data on regular basis (monthly, quarterly or annually) and have to respect a deadline for the data transmission. Depending on the data provider and the data set, the files contain from dozens of records to dozens of thousands.

 **A status report presenting the data delivery dates, the number of records received and highlighting the late data providers is maintained.**

In case of non-respect of the deadline, a reminder might be sent to the data provider.

3.2 Detection of the file format


The files received can be either flat (CSV or TXT) or Ms Excel. The separator used can be anyone (comma, semicolon, tab, etc.). The decimal separator (“.” or “;”) can also differ depending on the data provider. A header or footer can be provided or not.

 **The format can change depending on the provider of the file but can also differ for a same provider depending on the period covered. It is also possible that the format is different depending on the dataset.**

All flat files are manually opened before any treatment in order to determine the correct format and the adequate procedure to be used.

3.3 Checking of the content of each field

There are two types of fields: dimensions and values. For each dimension, only a specific list of codes can be used in each respective field. Basically, the content of each field is compared to the list of the corresponding acceptable codes. The checks on the values are very simple in a first step: checks on negative values and on non-numerical values. Each record, for which an error is detected, is rejected and stored in a specific table for reporting, analysis and checking.

 **Depending on the errors obtained, the user can decide to apply corrections and integrate the corrected data. A reporting of the errors found and of the corrections applied is produced for distribution to the data providers and the different actors intervening during the data treatment.**

Once these basic checks of the content of the file finished, the data are integrated in a production database containing the raw data as received from the data providers minus some minor corrections.

OW2 BI Initiative deliverable	Title: End users' scenario for OS BI: Statistical Data Life cycle Management	January 29 th , 2009
Author: Artemis	Revision: Engineering	Page: 9/12

3.4 Production of quality reports

Once the data integrated in the production database, some quality reports are produced: inter-dataset and intra-dataset checks.

The inter-dataset checks consist in comparing the data between the different datasets provided by a data provider for a given period. For example, two datasets can contain the same information but not at the same level of aggregation. The data provided in both datasets have to be consistent when comparing them at the same aggregated level. Another example is to compare that the data sent in quarterly datasets are equivalent to the data provided in annual datasets.

The intra-dataset checks consist in analysing the consistency of the data within a dataset. Two important checks are the analysis of the time series and the mirror checks. The reports on time series present the data for several periods of time (quarters or years) and highlight the potential problematic data detected according to predefined statistical formulas and thresholds.

Example of report on time series:

NB: The figures presented are fictitious.

Time series quarterly data - Dataset A1 - Gross weight of goods (in tonnes)

Port Code	Port Name	Direction	2005		2006				2007			Difference 2007Q03 - 2006Q03	Growth 2007Q03 - 2006Q03
			Q03	Q04	Q01	Q02	Q03	Q04	Q01	Q02	Q03		
ES01ESGJ	Gijón	Inwards	4 452 067	4 804 908	4 727 743	4 283 440	4 559 647	5 019 362	3 789 601	4 714 724	5 593 319	1 033 672	22.67%
		Outwards	409 439	410 039	399 911	551 484	418 114	501 034	379 209	426 142	431 941	13 827	3.31%
		Total	4 861 506	5 214 947	5 127 654	4 834 924	4 977 761	5 520 396	4 168 810	5 140 866	6 025 260	1 047 499	21.04%
ES02ESBCN	Barcelona	Inwards	6 188 211	6 166 881	6 235 524	6 189 386	5 972 651	6 472 985	6 378 064	6 374 154	6 525 483	552 832	9.26%
		Outwards	3 138 410	2 998 210	3 099 339	3 443 312	3 477 398	3 376 019	3 524 731	3 803 539	3 880 425	403 027	11.59%
		Total	9 326 621	9 165 091	9 334 863	9 632 698	9 450 049	9 849 004	9 902 795	10 177 693	10 405 908	955 859	10.11%

The mirror reports evaluate if the data provided by the different data providers are consistent according to predefined thresholds.



The reports are identical for each set of data and are produced after the reception of each set of data files. They are distributed to the data providers and to the different actors intervening during the data treatment.

After the diffusion of the quality reports, it happens that the data are not considered as sufficiently reliable. Most often, the data provider sends a new set of data.

3.5 Treatment of revised data

Usually, all set of data is revised and sometimes only partial data. The reception of revised data can imply three situations:

- Records have been provided with different values compared to the previous sending,
- Records remain identical to the previous sending,
- New records have been provided



The existing records are deleted from the production database and stored in a specific table containing historical data. Then all new records are imported in the database passing through all validation rules described above.

OW2 BI Initiative deliverable	Title: End users' scenario for OS BI: Statistical Data Life cycle Management	January 29 th , 2009
Author: Artemis	Revision: Engineering	Page: 10/12

Data can be revised several times. In consequence, the table containing historical data contains several versions of a data set. Only the last data sent have to be considered for dissemination purposes. The different versions of the data are kept for comparison purpose. Once data provision finalised and quality checks validated, the data are ready to be disseminated.

3.6 Data dissemination

The dissemination of the data can take several forms:

- Visualisation of the data in the production database and simple calculations/aggregations. Several reports are produced on regular or ad-hoc basis where the year, the countries or any other parameter for which the data have to be presented can be chosen.



Usually, users creates simple reports with the minimal information needed and exports the data in a spreadsheet-application in order to apply specific calculations on the individual figures presented in the report.

- Production of publications (leaflets, brochures, pocketbooks, etc...) using predefined templates on regular basis (yearly or quarterly), containing graphs, tables or maps presenting figures and text commenting on the figures. In most of the cases, these publications are produced in PDF format and disseminated on the web. Moreover, some of them have also to be printed and bounded.



“Pixel-perfect” quality is required for these publications

Several drafts of the publications are transmitted to data providers and to other different actors for control and validation before the final dissemination.

OW2 BI Initiative deliverable	Title: End users' scenario for OS BI: Statistical Data Life cycle Management	January 29 th , 2009
Author: Artemis	Revision: Engineering	Page: 11/12

Conclusions

This document investigates the Statistical Data Life cycle Management as a prominent end users' scenario for OS BI.

The needs of the different actors intervening in the statistical data life cycle are investigated and the description of a statistical DLM scenario is given.

A next version of the document will present how current OS BI tools already answer to user needs and will highlight the existing gaps. Finally we'll propose how this gap can be fulfilled in a two steps scenario: a mid-term goal in order to strengthen the current OS BI platform and tools to support current end users' needs, and a long term goal to envision how the second generation of business intelligence tools could include a specific support to Statistical DLM.

OW2 BI Initiative deliverable	Title: End users' scenario for OS BI: Statistical Data Life cycle Management	January 29 th , 2009
Author: Artemis	Revision: Engineering	Page: 12/12