Statistics Catalonia: towards total integration of statistics production

F. Udina[1], C. Rovira[2], J. Sort[3], M. Perelló[4]

[1] Statistical Institute of Catalonia (Idescat), Barcelona, Spain. fudina@idescat.cat
[2 3 4] Statistical Institute of Catalonia (Idescat), Barcelona, Spain

Abstract
We describe the ongoing process for modernization of statistical production and dissemination initiated three years ago at our institute. It is a two track process: on the business track, we proceed first to analyze and document all the statistical processes to redesign and integrate them, using three basic statistical registers (population, entities, territory) into the Integrated System of Statistical Information. On the technological track, we redesign our information management scheme based on four conceptual stages: Data capture – Integration – Stat product generation – Data access.

Keywords: (1-5 words), GSBPM; integrated statistical production processes; modernizing the production and dissemination of official statistics; statistical and admin data for research.

1. Introduction

The purpose of official statistics is to provide high quality statistics for a society to enable the improvement of public policies and the general good. A modern approach requires the integration of information from many sources, the high quality of production processes and the use of multiple channels of dissemination to reach a wide variety of users, requiring constant and strict control of data security and confidentiality.

Over the last three years, Idescat has conducted a two-track innovation process to achieve these goals: on the business track, the focus has been on the implementation of a documentation production metadata system based on GSBPM (Qualitas project) which guides the modernisation of statistical processes; on the technological track, the focus has been on the implementation of a new information management scheme (Cerdà Platform) based on four conceptual stages: Data capture – Integration – Stat product generation – Data access.

The Cerdà Platform is entering the final phases of implementation and is allowing the integration of the incoming data (validated in the first stage) by means of coding information on people, economic entities and territory using the three core statistical registers. Thus, every piece of integrated information is available to statisticians in all the areas of the Institute. Using an anonymization process the combined information is made available to different levels of users.

The Qualitas project provides methodological guidance, quality control and metadata production to implement the desired statistical products. Starting with some proofs of concept, all the current statistical production is being moved to the new model.

Among other goals, we aim to provide statistical data from several origins (admin data, survey data, big data, etc.) to researchers in social sciences. Taking care of data security and confidentiality issues is crucial here, as is developing safe protocols to ensure that good researchers have high quality data to develop safe projects.

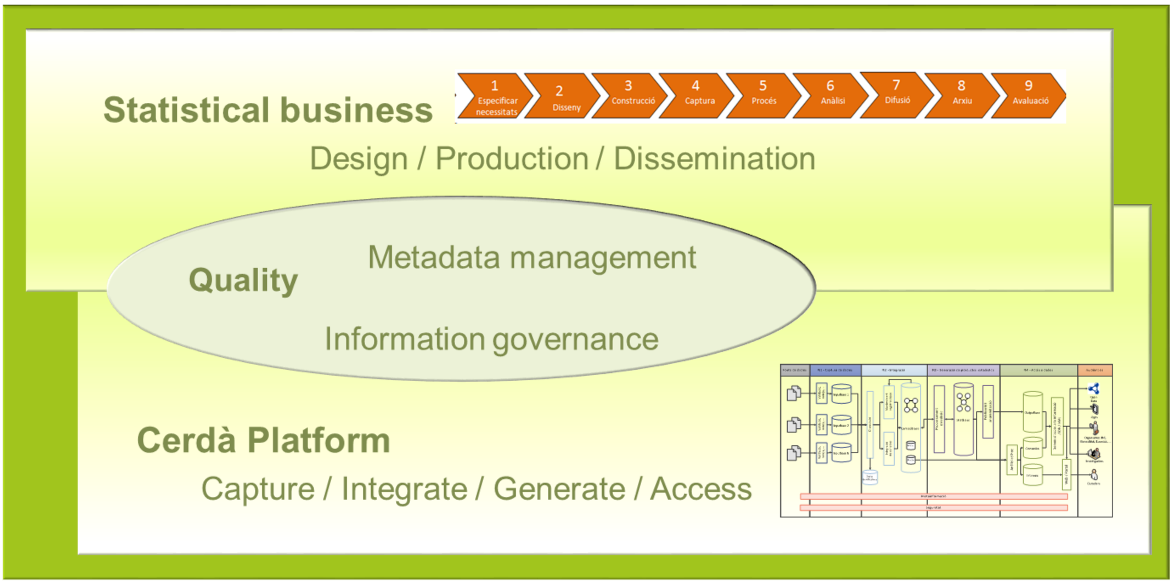


Fig. 1.- The Integrated Statstical Information System

2. The Integrated Statistical Information System

One major aspect of modernizing statistical production is information integration allowing the efficient reuse of information and combining different sources of data. The old *stove pipe model*, whereby each statistical operation is performed without any connexion to any other (Figure 2 offers a good visualization of the old model) needs to be replaced by a model in which many sources of data can help to build more complex statistical models.

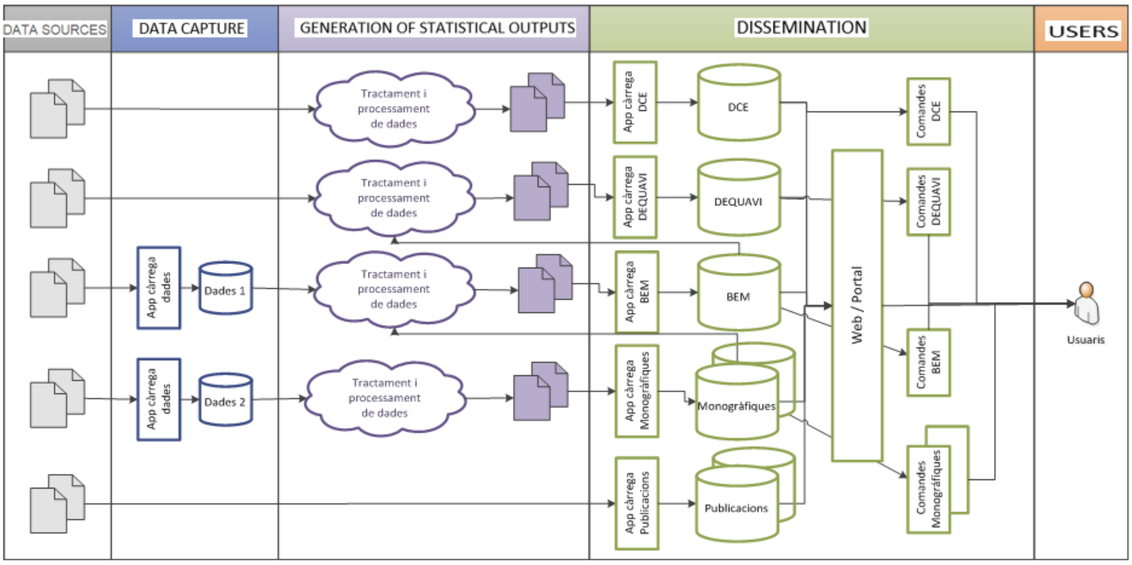


Fig. 2.- Information management: scheme previous to the project.

Our approach to information integration is based on (Wallgren and Wallgren, 2014) without seeking full implementation of their model. We use administrative information to build the core of the system: the three basic statistical registers of population, business and territory. The main role of these registers is to replace any information on people, business (in a broad sense of the concept) and territory with coded information. Each individual will have a code to allow linkage of information and to avoid working with identifiable fields. Each business (i. e. any organization with economic activity, including firms, local units, public organisms, non-profit entities, schools, universities, etc.) will also have its own code, while any territorial information, street addresses, cities or municipalities will be coded using the territory register. Any identifying data will be kept apart and safely protected by restricted and controlled access.

With the above coding of the main fields, linking the information should be easy when a statistical operation needs to use information from several sources,.

In this way, we want to be prepared to use several sources of information: primarily administrative data. Then, when necessary, we will also process data from surveys, other administrations or firms, the network and other big data sources, etc.

In the next two sections, we describe the two tracks mentioned in the introduction: the technological track to support the information flow and the business track to analyse, document, and transform the statistical processes.

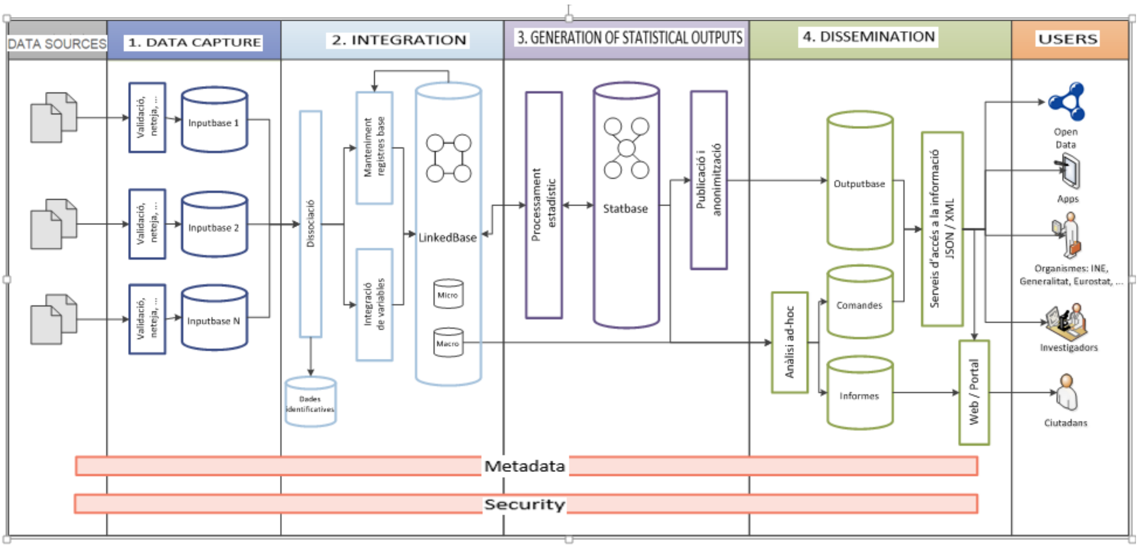


Fig. 3 .- The new information management scheme: The Cerdà Platform.

3. The technological platform

To support an integrated system of information we need to design and implement a new technological platform, which we have called the Cerdà Platform. The scheme in Figure 3 may help to understand the changes, especially if it is compared to the one in Figure 2.

As shown in the figure, the new platform has four levels for transforming the data sources into useful information for final users: Data capture, Integration, Generation of statistical output and Dissemination. We will briefly describe each of these steps.

3.1. Level 1. Data capture

On this level, data from several sources (the administration, surveys, private sources, etc.) are checked for integrity and validated.

3.2. Level 2. Integration

The first step on this level is to replace any identifying information with its codes according to the basic statistical registers. The identifying information is kept apart in a safe place. A series of data quality processes is then applied: data fusion or linkage when necessary, detection of duplicates, imputation of missing data, small area estimation or other statistical methods. We call the final outcome of these operations *LinkedBase*.

3.3. Level 3. Statistical product generation

On this level the statistical staff select the required data from among those available in the *LinkedBase* to produce the desired statistical output. Specialized software may be used here: e. g. SAS, SPSS, R, JDemetra+, etc. Aggregation in time or space is performed and the rules for preserving confidentiality are decided at this point. The resulting *StatBase* may also be used as input data for subsequent operations.

3.4. Level 4. Dissemination

Statistical products are stored in the *OutputBase* in multi-dimensional cubes with confidentiality and quality guarantees, ready for final user access. All the cubes contain at least one variable, together with time and territory dimensions. Some kind of *business intelligence* tool will allow users to extract the information.

API tools and open format JSON-Stat are used to access the selected information,. Several channels are available for different kind of users (web, open data, special apps, secure access channels, etc.).

All the levels and data communication across the Cerdà Platform are governed by metadata and safety protocols to ensure confidentiality and data security. State-of-art tools are used on each level, for example to extract-transform-load pieces of data, to select sets of cases and variables by means of BI tools, etc.

4. Statistical processes analysis and integration: The Qualitas project

The Qualitas project has two phases: firstly, extensive and standardized documentation regarding the statistical processes currently active at the Institute is collected. In the second phase, the above processes need to be analyzed in order to improve efficiency and to proceed to integrate them into the new paradigm of production.

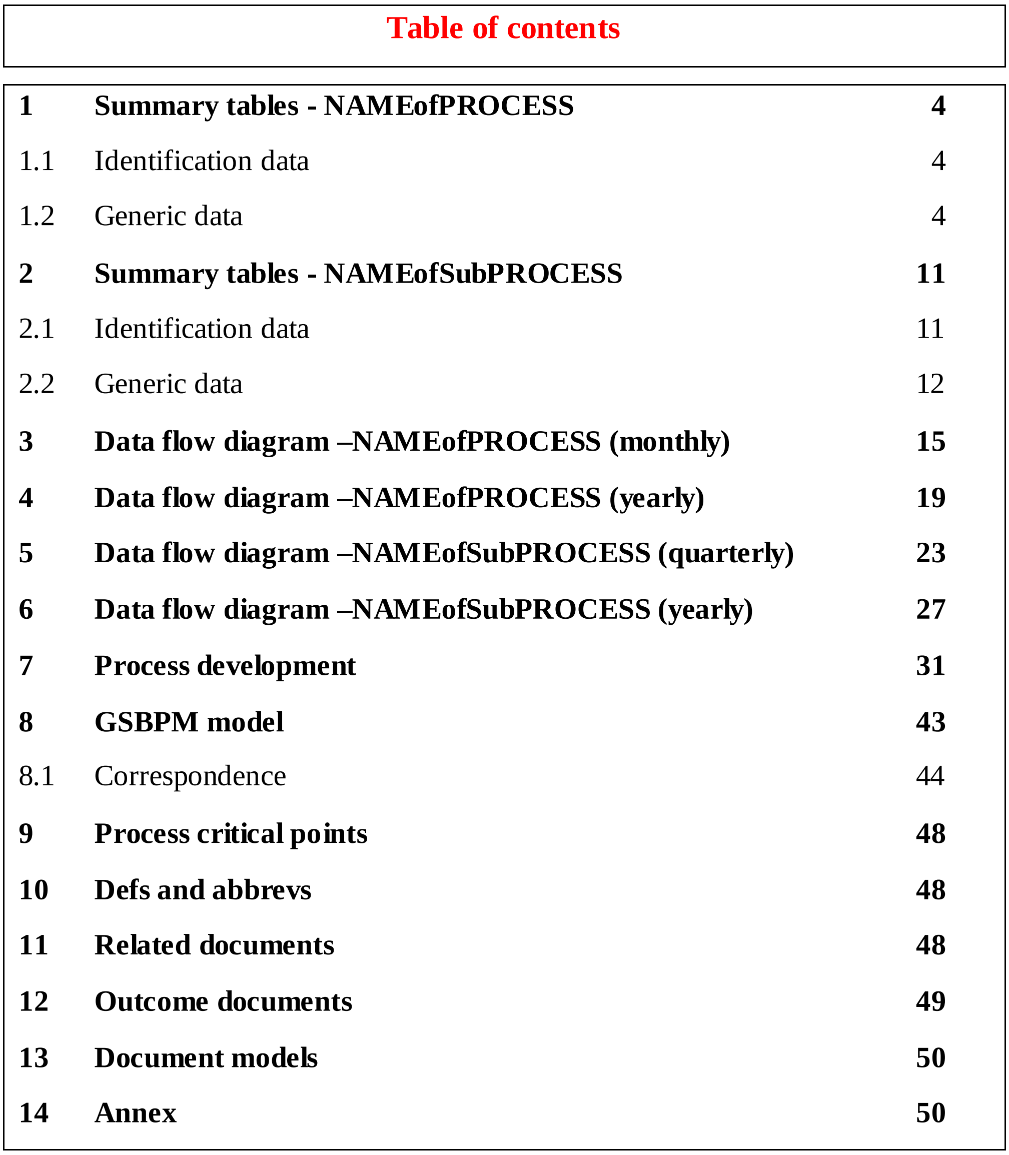


Table of contents

1	**Summary tables - NAMEofPROCESS**	**4**
1.1	Identification data	4
1.2	Generic data	4
2	**Summary tables - NAMEofSubPROCESS**	**11**
2.1	Identification data	11
2.2	Generic data	12
3	**Data flow diagram –NAMEofPROCESS (monthly)**	**15**
4	**Data flow diagram –NAMEofPROCESS (yearly)**	**19**
5	**Data flow diagram –NAMEofSubPROCESS (quarterly)**	**23**
6	**Data flow diagram –NAMEofSubPROCESS (yearly)**	**27**
7	**Process development**	**31**
8	**GSBPM model**	**43**
8.1	Correspondence	44
9	**Process critical points**	**48**
10	**Defs and abbrevs**	**48**
11	**Related documents**	**48**
12	**Outcome documents**	**49**
13	**Document models**	**50**
14	**Annex**	**50**

Fig. 4. An example of process documentation table of contents

4.1. Documentation phase

The information collected for each document is summarized in Figure 4. Information on the persons responsible for the process and each of its phases has been recorded in the *Identification Data* section. The *Generic Data* section includes legislation and other applicable rules and a summarized description of the process and its phases. In some cases, a sub-process is identified and described. The data and information flow to support the process have also been carefully documented (see Figure 5). A more detailed description is provided in the *Process Development* section. Section 8 is devoted to the correspondence of the process analysis to the Generic Statistical Business Process Model (UNECE 2013), the current standard for official statistics, adopted by UNECE, Eurostat and other main statistical offices (see Figure 6). In our documentation we attempt to relate the phases and steps described to the ones regarded as standard by the GSBPM.

A key issue on quality for statistical processes (see Ehling et al, 2007) is identification of process variables. We call it *critical points of the process* and Section 9 is devoted to it. The rest of the documentation includes references to related documents and other useful information.

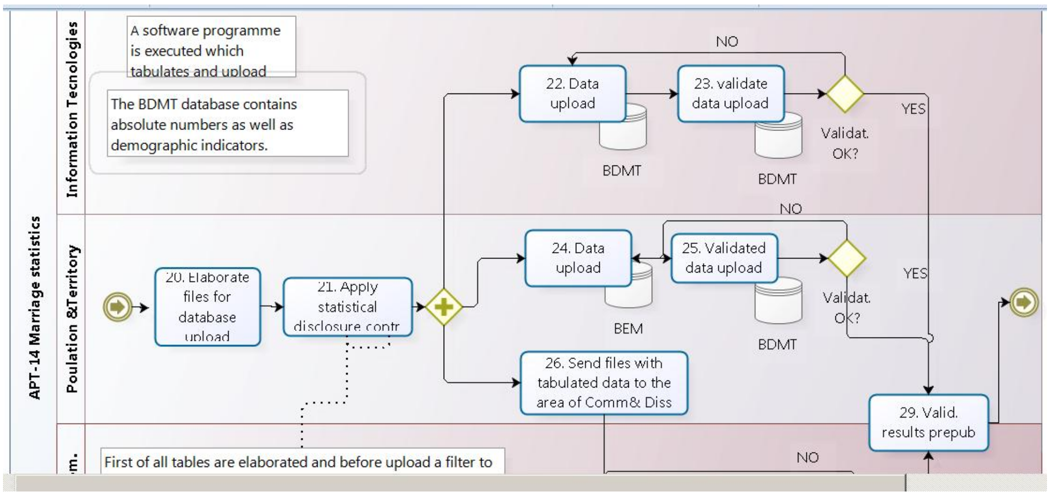

APT-14 Marriage statistics
Information Tecnologies
Poulation &Territory
m.
A software programme is executed which tabulates and upload
The BDMT database contains absolute numbers as well as demographic indicators.
NO
22. Data upload
23. validate data upload
BDMT
BDMT
Validat. OK?
YES
NO
20. Elaborate files for database upload
21. Apply statistical disclosure contr
24. Data upload
25. Validated data upload
BEM
BDMT
Validat. OK?
YES
26. Send files with tabulated data to the area of Comm& Diss
29. Valid. results prepub
First of all tables are elaborated and before upload a filter to
NO

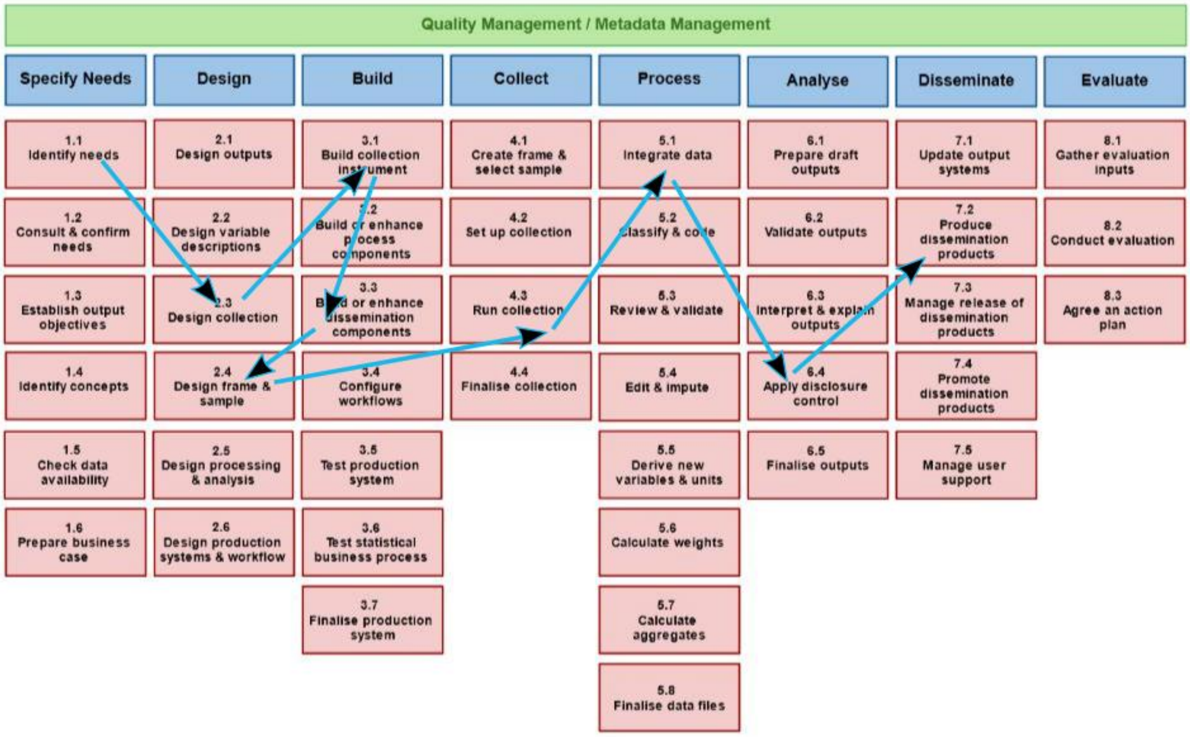
Fig. 5. Data flow example, specifying departments in charge of each step.

Until now a total of 73 production processes, many of them with sub-processes, have been studied and documented. They correspond to the four Idescat areas and are mainly devoted to productive rather than instrumental or dissemination processes. Business & Occupation, Economy & Society and Population & Territory account for the bulk of the processes (62). A total of 70 meetings were held and 48 people of Idescat staff were involved in discussing and compiling all the information.

4.2 Results and further work

The main result of Qualitas is the complete documentation of the tasks, which is important because it makes the processes reproducible regardless of the persons/areas in charge. Some other effects of the work performed are the discovery of duplicities and possible shortcuts in the processes, the identification of important files and programs, having an updated list of the tools used in the processes, etc.

One significant step forward is building a document management tool to keep the Qualitas documentation alive and up-to-date. A metadata management system will be more easily built in the near future by means of this tool. This metadata system is to evolve for becoming the common backbone to the Cerdà Platform.



Quality Management / Metadata Management

Specify Needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Build collection instrument	4.1 Create frame & select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult & confirm needs	2.2 Design variable descriptions	3.2 Build or enhance process components	4.2 Set up collection	5.2 Classify & code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collection	3.3 Build or enhance dissemination components	4.3 Run collection	5.3 Review & validate	6.3 Interpret & explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
1.4 Identify concepts	2.4 Design frame & sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit & impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing & analysis	3.5 Test production system		5.5 Derive new variables & units	6.5 Finalise outputs	7.5 Manage user support	
1.6 Prepare business case	2.6 Design production systems & workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production system		5.7 Calculate aggregates			
				5.8 Finalise data files			

Fig. 6. Correspondence to GSBPM model

The second phase of the Qualitas program is to study, analyze and convert, one by one, the statistical processes to the new paradigm whereby the statistical output is produced using all the related available information, starting from the core of the statistical registers, and making the outcomes of each process available for subsequent statistical products. This will enable us to achieve the relevant aspects according (ESS 2011): sound methodology, appropriate statistical procedures, nonexcessive burden on respondents and cost effectiveness.

5. A final note on confidentiality and usefulness of information

As mentioned above, we care about the diversity of the final users of our official statistics outcome. We believe that the main purpose of official statistics is to help in the decision making, measurement and improvement of public policies. Thus, administration managers and social science researchers are among our users, as are other economic and social actors, the general public and other social groups which are involved in controlling public activity.

Having this diversity of potential users requires careful deliberation with regard to how the data should be delivered to each group of users. Open data is an excellent asset for all citizens and social movements with tools to control public activity, but they have a severe limitation: concern for privacy. Delivering rich multi-dimensional data is incompatible with confidentiality; the richer a dataset is, the easier it is to re-identify any subject of the information. Between the Integrated System, which can only be used by the internal staff of the Institute, and Open Data, that is for the general public, we are considering different levels of aggregated or anonymized information as shown in Figure 7: confidential microdata for scientific purposes, aggregated data available via an extranet for administrative units and controlled access, Data on Demand and, finally, Open Data.

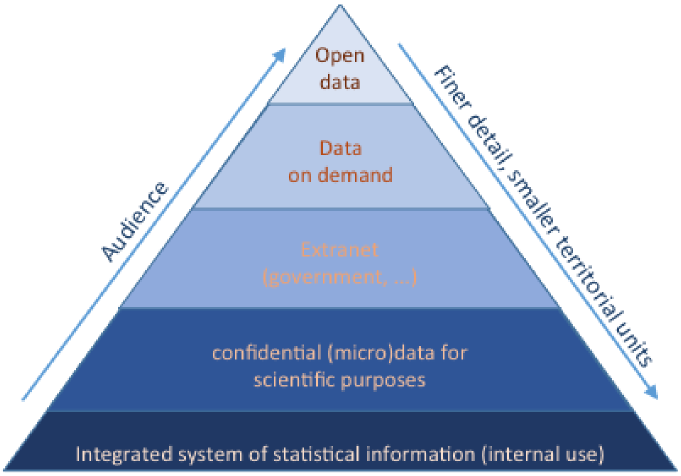

Open data
Data on demand
Extranet (government, ...)
confidential (micro)data for scientific purposes
Integrated system of statistical information (internal use)
Audience
Finer detail, smaller territorial units

Fig. 7. The data dissemination triangle.

6. References

Ehling, M. and Körner, T. (eds.) (2007). Handbook on Data Quality Assessment Methods and Tools. Eurostat.

European Statistical System (2011) European Statistics Code of Practice - revised edition 2011

UNECE (2013) The Generic Business Process Model v 5.0. http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0 (as found in April 2016)

Wallgren A. and Wallgren B. (2014), Register-based statistics. Statistical Method for Administrative Data, 2nd edition. J. Wiley.