STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Fifty-third plenary session
(Geneva, 13-15 June 2005)

DATA COLLECTION STRATEGY AND INTEGRATED SOLUTIONS

Supporting paper submitted by ISTAT, Italy∗

ABSTRACT

The relationships between data respondents and data collectors in the work of CODACMOS project are approached under a given framework of strategy for an automated data collection. First of all, from the integration of primary and secondary data collection point of view, the data collection strategy involves the integration of primary and secondary data; systems and levels; the integration can increase the use of secondary data, lower the burden and improve data quality. But, primary and secondary integration has specific issues: the data are collected by different organizations, they may be collected by different modes and at different points in time and the related metadata may differ.

The integration of primary and secondary data collection studied under CODACMOS project has been carried out in two axes: in types and in levels. The first one dealt with and demonstrated the feasibility of the integration of data from different sources, integration of primary and secondary systems and the integration of existing data (simplification of tasks of reporting and confirmation of data). In addition to, the integration is analysed not only as a technological process only, but also a conceptual and organisational process. As a process, it includes at least three phases: a phase of integration to level of data, a phase of integration to level of process and finally a phase of integration at state level.

∗ Prepared by A. Sorce and P. Dishnica.

GE.05-

In that context, from the metadata point of view, CODACMOS has found out that the metadata used under integration and data sharing process and exchange should be captured (exchanged) along the collection of data (at least for statistical process) in order to guarantee their re-use in the following stages of the process (processing and dissemination) rather than inventing from the scratch every time some "explanatory information". Their connection is strong: the metadata are shared and reused across different data sets, in particular during the integration stage (the exchange and sharing of metadata coming from multiple sources and repositories). The metadata of metadata repositories maintained within the same organisation that are matter of exchange (or data collection), by their re-use and sharing, have to associate the relative data up to the final destination, the users. In such a situation, i.e. in order to ensure the coherence and data quality between various data collectors, these users have the right to ask for more background information on the data already available or shared (how the terms are defined, how they relate to neighbouring information in the same environment and how the data are collected).

The scope of this paper is to present some of the most agreed and feasible results of the work carried out during CODACMOS project dealing with the integration and data sharing issues as well as some problems related on, in particular for a National Statistical Institute.

INTRODUCTION

1.   To facilitate an effective electronic data collection process minimizing the administrative burden for the respondents for a National Statistical Institute is not at all an easy task. It depends from many factors and existing environments inside and outside the organisation.

2.   Production services of NSIs are still more data collection oriented rather than user/customer oriented, the proposals made by CODACMOS project tend to balance better the passage from the traditional to the modern system of information and organisation. That means that before taking into consideration any of the proposals made, some revision, rationalisation of the existing situation and the state of art is necessary. In particular, statistical information system content, state of art and management should be matter of review and analysis.

3.   Some proposals request only an optimisation process. Other are more method and tool oriented: it is a matter of fact that more common understanding on definitions, methods, models and tools and future vision are necessary. Accepted glossaries, guidelines, users' guides and technical recommendations are needed too. How helpful are the existing systems of metadata for the actual situation of NSIs and for the future ? Which are the most relevant requirements of internal users of European and National statistical system and the external ones ? How the existing IT infrastructure of NSIs should reflect the requirements for integration, for standards and for more flexibility ? Furthermore, it is obvious that the definition of an Automated Data Collection and metadata strategy has its significant implications on the IT architecture.

4.   Organisational and management issues within NSIs as well as the ones dealing with a better coordination and collaboration with non statistical data collectors in the framework of eGovernment make a significant sense when integration, sharing, standardisation and

harmonisation become a priority not only at organisation level but national level too.

5.   The vision of the automated data collection and metadata strategy is for most of NSIs to have their rightful places in the international community of official statistics, subscribing to international standards for quality and practice (described through the collection, harmonisation, processing and analysis and dissemination of data). If the vision and strategies proposed under CODACMOS project will be matter of further discussion and research, a re-engineering of key statistical systems based on a modernisation programme could be the medium (long)-term framework initiative to be launched by NSIs.

6.   Technical meetings and discussions have to be performed with EUROSTAT, OECD, UN and NSIs experts involved in various statistical activities in order to concretise the feasibility of CODACMOS results and their exploitation.

## CODACMOS DATA COLLECTION STRATEGY UNDER EGOVERNMENT SCENARIO

Moving towards new scenarios of data collection and exchange

7.   The starting point of the data collection process and exchange that CODACMOS has identified is the existence of the following practices and scenarios:

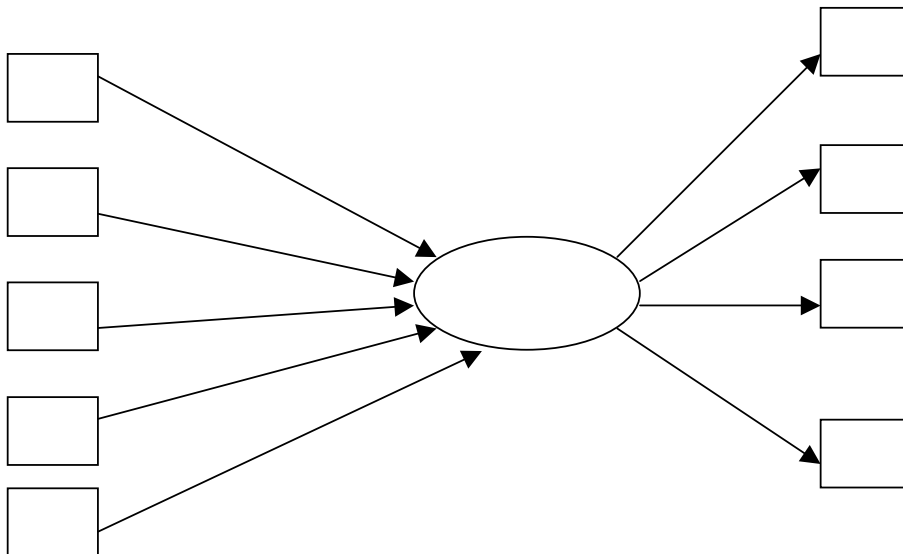(model M0): N data respondents   -   N data collectors;
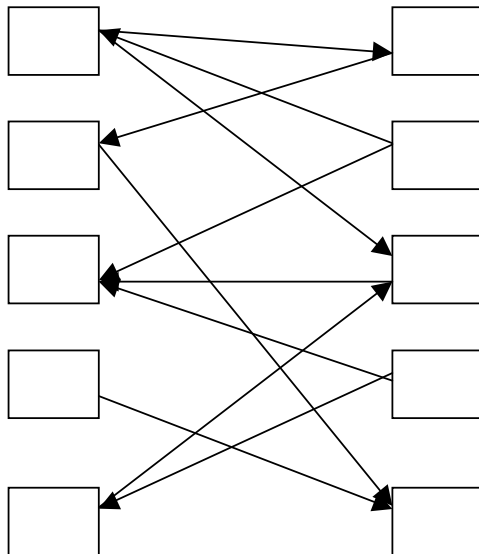
(model M1): N data respondents   -   1 data collector;
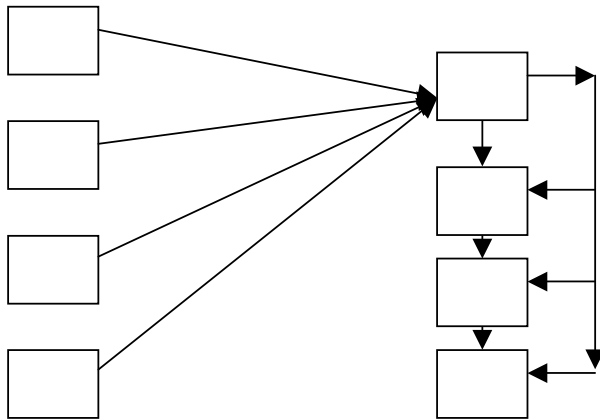
(model M2): N data collectors (data respondents) - N International Organisations (data collectors)
          (OECD data sharing model).

8.   This latter deals with the exchange of macrodata, but it can be applied for microdata in the context of data collectors and data respondents.

9.   The M1 and M2 features relates to:
•   The dimension of extension or the size: from this point of view there can be identified three kinds of levels of organisations involved in the process of data collection and exchange: local, national and international organisations.
•   From the technological point of view: the Finnish model uses the file transfers (both

FTP and HTTPS) and WEB forms and secure eMail while the data sharing SDMX model is based on WEB technologies. The difference exists on the technology of data transmission. In the first case the data and the metadata can be sent separately, in the second case the metadata are encapsulated within the data.

- From the micro/macro data point of view: the definition micro and macro data which are the basis of these two models became relative. It depends from the operational environment, i.e. the data that are defined as macrodata for the model M1; the M2 model calls these data as microdata.

10.   In order to link these two models, some conceptual and strategic (eGovernment) common understanding is needed. The up to date technologies supply solutions of any question. The standardisation of the definition of metadata becomes essential. Approaches and experiences on the matter exists in the practice. The metadata in this case are associated (encapsulated) to the data instead of the other experiences of the actual situation where the data and the metadata are separated.

Organizational issues come first. Connection of statistical data collection process to eGovernment program

11.   E-Government is a very widely used term. It has many meanings and implementations also. Progress has been significant over the past years.
Data Collection is one of the key parts of the e-Government initiatives. All European e-Government portals already offer data collection and automated reporting facilities. The starting point of these initiatives has been the need to help enterprises to fulfil their administrative duties.

12.   An effective implementation of electronic (automated) data collection within e-Government means easily accessible data for the data collectors, whereas the data provider has the minimum administrative burden. This means reduction of duplicated data collection, minimising the collection of new data, rational collection methods and optimisation of the collection processes. The natural way to do that is the national, cross-sector, co-operation plan that is the key part of the e-Government implementation.

13.   The rationales of the e-Government data collection are apparent:
- Use and utilize base registers,
- Use basic numberings systems (PIN, BIN, RIN, ..)
- Use and develop data element and data record standards
- Use and develop standard interfaces
- Use and develop data delivery channels and services
- Build seamless data transfer chains
- Look at the data collection process from the reporters' point of view (because they do the biggest job)
- Use and develop key components (XML schemas, record structures, syntax checking services, web services, forms libraries, file transfer services, authentication services, digital signature architectures, payment services, software architectures, common portals and collaborative data systems)
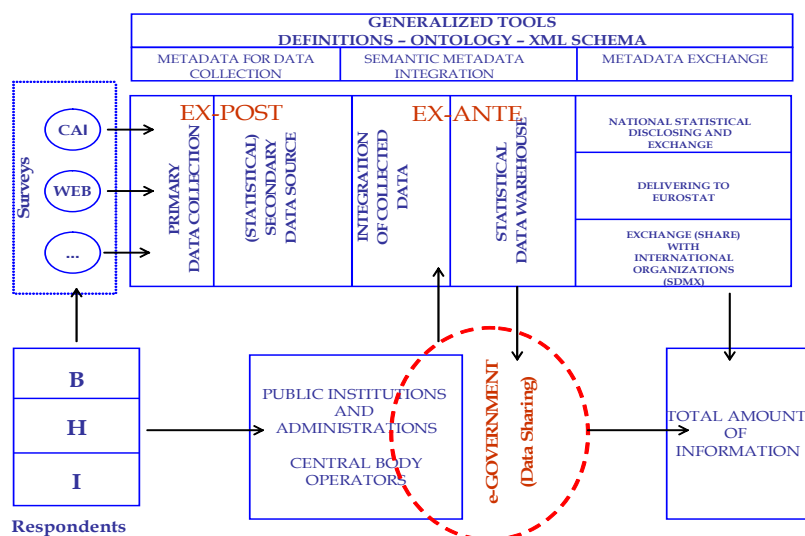- Co-operate at least on three levels:

- National cross-sector level (having common customers).
- International intra-branch level (having common challenges).
- Technology levels (having common facilities and visions).
- Learn and follow other's experiences but take your own step having your own direction.

14. In the planning of e-Government activities it is vital to find a common goal for the participating governmental institutions. E-Government cannot be realized in a vacuum. Its success is closely linked to fundamental change of public governance and administration. E-Government should concern the state or European Union as such, but on practical point of view, the starting point should be finding the most interesting partners to get the first results. Prerequisite for close partnership can be e.g. common data or work flow, usage of similar architecture or tools. In Finland, Statistics Finland gets 90% of data from other data collectors, one of these is tax administration. Another example is identification: it is very reasonable to co-operate and arrange joint-service, because all governmental agencies need identification on their electronic services.

15. Understanding between the partners is the first step: what is their core "business", competence, needs etc.. It is important to discover common denominators and start from the point that is possible from each partner's point of view. However, it is more than probable that changes or revisions have to be made in the processes or workflows to find the common denominator. Interoperability and standardization feasibility are two very important issues identified for achieving a common understanding both on the technical level and on the political/administrative level. The biggest barriers to e-Government are often in the minds of the people. Planning and implementing must have approval from the highest level in each organization. Political and administrative reality should keep in mind. E-Government is not just public: private data collectors, intermediaries, banks, software vendors should be involved in e-Government. Use the experiences and skills of the private sector, rather than reinventing existing solutions. A danger of implementing e-Government to streamline the data collection process is bureaucracy. Of course, when co-operating and collaborating the sharing of information is vital but it should not mean more bureaucracy. It is avoidable if goals are set clearly and goals are achievable. Another major factor is that the partners should trust each other in their co-operation. Producing huge amounts of paper, empty meetings or intentional bureaucracy does not compensate lack of trust.

16. CODACMOS has developed a general scenario for Data Collection and dissemination process framed to an e-Government experience. The following figure shows how the statistical process and the e-Government are connected to each other. It could be as the starting point for the definition of the strategy of data and metadata collection and exchange. The rationale of this general scenario comes from the SISSIEI model developed at ISTAT.

**General Scenario of Data and Metadata for Data Collection in a
NATIONAL STATISTICAL INSTITUTE**

GENERALIZED TOOLS DEFINITIONS – ONTOLOGY – XML SCHEMA		
METADATA FOR DATA COLLECTION	SEMANTIC METADATA INTEGRATION	METADATA EXCHANGE

EX-POST   EX-ANTE

Surveys
- CAI
- WEB
- ...

PRIMARY DATA COLLECTION

(STATISTICAL) SECONDARY DATA SOURCE

INTEGRATION OF COLLECTED DATA

STATISTICAL DATA WAREHOUSE

NATIONAL STATISTICAL DISCLOSING AND EXCHANGE

DELIVERING TO EUROSTAT

EXCHANGE (SHARE) WITH INTERNATIONAL ORGANIZATIONS (SDMX)

B
H
I

Respondents

PUBLIC INSTITUTIONS AND ADMINISTRATIONS
CENTRAL BODY OPERATORS

e-GOVERNMENT (Data Sharing)

TOTAL AMOUNT OF INFORMATION

17.	If the main data collectors at national level will include the NSI to collaborate according to a result of co-operation agreement (e-Government programme), it means that the information could be collected once and the data collector is well identified. So, the technical aspect of data collection becomes a strategic problem closely linked to the context of e-Government.
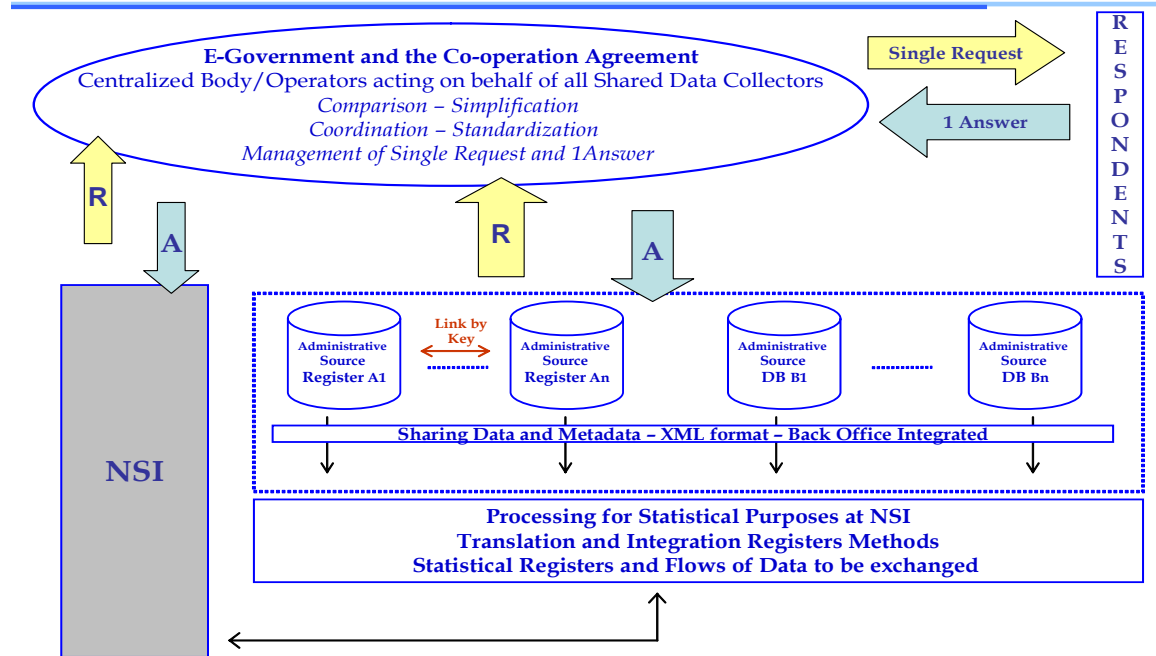
18.	In such a context, two scenarios on the organizational level are developed as in the following:

Scenario 1

19.	The NSIs and other shared data collectors are framed under the eGovernment project and a common co-operation agreement, where **a centralised body** or operators act on behalf of all shared data collectors. Four main processes are foreseen: comparison, simplification, coordination and standardisation. The respondents receive the single request (due to this four processes) and answer only once (1 Answer).
The data collected by non-statistical data collectors are used for statistical purposes on the basis of the processes of translation, register integration and flows of data to be exchanged. The centralised body is located at national level (outside NSI). But, of course, in general sense, the functions of coordination and standardisation are at least part of the main activities of a statistical organisation.

Data Diagram for NSIs: Organization Level (Scenario n. 1)

E-Government and the Co-operation Agreement
Centralized Body/Operators acting on behalf of all Shared Data Collectors
Comparison – Simplification
Coordination – Standardization
Management of Single Request and 1Answer

Single Request

1 Answer

RESPONDENTS

R

A

R

A

NSI

Link by Key

Administrative Source Register A1

Administrative Source Register An

Administrative Source DB B1

Administrative Source DB Bn

Sharing Data and Metadata – XML format – Back Office Integrated

Processing for Statistical Purposes at NSI
Translation and Integration Registers Methods
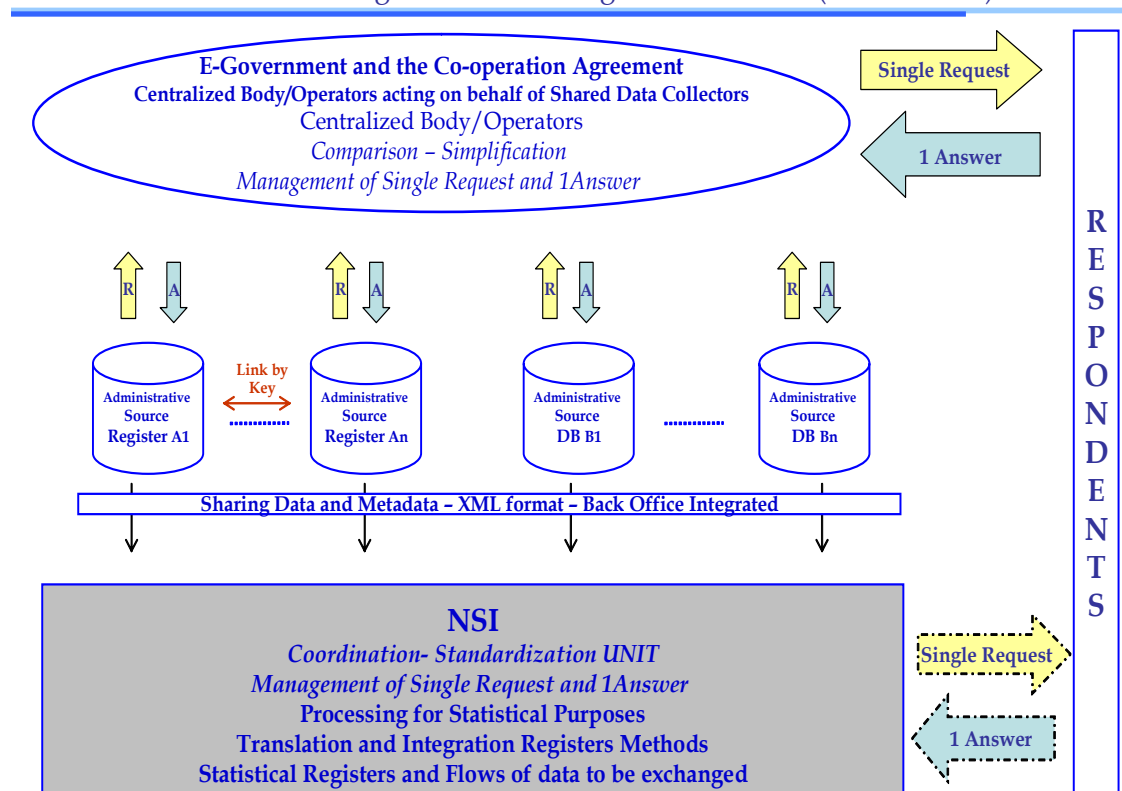Statistical Registers and Flows of Data to be exchanged

Scenario 2

20.   In this scenario, the activities of comparison-simplification are separated to those of coordination and standardisation. This latter is organised as a unit at NSI level.[1] As soon as the confrontation with the existing shared data and metadata is achieved and the single request has been formulated, the NSIs could start a primary data collection process at respondent level, guaranteeing that to this request for information there are no other data collectors that repeat the same request. As to the previous scenario, the data collected by non-statistical data collectors are used for statistical purposes on the basis of the processes of translation, register integration and flows of data to be exchanged.

Data Diagram for NSIs: Organization Level (Scenario n. 2)

E-Government and the Co-operation Agreement
Centralized Body/Operators acting on behalf of Shared Data Collectors
Centralized Body/Operators
Comparison – Simplification
Management of Single Request and 1Answer

Single Request

1 Answer

R
E
S
P
O
N
D
E
N
T
S

R A   R A   R A   R A

Administrative Source Register A1   *Link by Key*   Administrative Source Register An   Administrative Source DB B1   Administrative Source DB Bn

Sharing Data and Metadata – XML format – Back Office Integrated

NSI
Coordination- Standardization UNIT
Management of Single Request and 1Answer
Processing for Statistical Purposes
Translation and Integration Registers Methods
Statistical Registers and Flows of data to be exchanged

Single Request

1 Answer

What is missing actually?

21. Brainstorming, cooperation and common data collection framework if the integration (at state level) is considered as well as a high priority for improving the quality of official statistics, lowering the burden and the costs for both data collectors and providers. "The way for reaching this goal is to improve the efficiency of data collection and integrate the data exchange between various data collectors" (CODACMOS Project).

22. Even a statistical system has enough technical capacity, the essential problem is about governance. So, it is important to discuss technical issues but it is even more important to develop a vision of the new system. This is what Codacmos has contributed.

23. Recent technical developments could give a prominent role to the NSIs in providing web services to other administrations in the eGovernment context and this can give a role to NSIs also in terms of global governance of national statistical systems.

24. For example, in Canada, Statistics Canada has a particular institutional set up since there is a long tradition in Statistics Canada to have a leadership role in this respect. How much NSIs can learn from such experience ?

25. In the following, some common elements and goals between Statistical process and e-

Government process are identified. They might be useful for NSIs to feed up this process in their countries.

Description	Statistical process	e-Government process
Electronic (digital) data collection/ production and data dissemination	(elementary) Statistical and Administrative data collected for statistical purposes	Administrative data (collected, produced and distributed) contained in data bases (developed and maintained)
Integration/sharing of information Model(s)	Primary and secondary sources of data Integration of data sources and registers Collect "only once" approach	Integration of systems. The production of information takes place in a network (multitude of back offices). Whilst production takes place in a multitude of networked back offices – often far from the end-user – the distribution of these services may take place either virtually or physically in front offices.
Timeliness and Costs	Quicker and lower	Very low – cheaper
Accessibility	Wider (data protection)	total
Quality	Data quality	Data quality, service quality
Respondents:	Individuals, households, businesses, intermediaries, government bodies (national/local), focus: respondents first	Individuals, businesses, government bodies (national/local) Focus: citizens (individuals) first
Lowering of burden	Statistical burden on respondents	Administrative burden Ensuring multi-channel interaction system with administrations, to avoid relying exclusively on internet
Interdependence for strategy design (cooperation between actors)	Statistical data collectors, ADC strategy design	Non-statistical data collectors: govt agencies (and other users, universities, stakeholders, SW companies, etc.) - Part of e-Government programme
Technology and standard problems	Stand alone off line solutions, standard data capture software, self-interviewing software, commercial packages, www- forms, e-mail, electronic storage of databases, PKI & encryption, common standards	File transfer- when transferring large amount of information; request/reply to handle queries typically involving smaller amount of information, self-interviewing. e-mail, electronic storage of databases, PKI & encryption, common standards

26.   The statistical data collection means:
•   Direct and specific primary data are used

- Using the secondary sources (optimization)
- Better utilization of the already collected (and handled) data
- More efficient statistical production processes
- Better response to the needs of statistical data customers
- New data delivery strategies: publicity, sharing, accessibility and understandability

27.   Some lessons learnt in relation to the common elements between e-Government and Statistics request could be listed as in the following:
- Standardization (metadata) to be maintained.
- Common framework in terms of cohesion and understanding between partners.
- Integration (in particular statistical integration, metadata integration) and storage of public metadata is possible (public repository can avoid duplications).
- Quality can be improved.
- Security to be maintained and balanced (there are differences).
- Models/scenarios (data sharing projects are possible), but establishing uniform processes or procedures is not easy.
- Technology (gaining expertise with new technologies).
- Cooperation is possible, but a coordination of efforts within and between organizations has to be improved.

28.   In relation to the Codacmos problem "How to improve processes between the data collectors and the data respondents with the primary focus on the respondent needs, in order to improve the data quality and render easier and more efficient the task of electronic responding; to obey to the rules of confidentiality and to simplify the information asked", what Codacmos has developed is the new approach on data collection and exchange process: it is a matter of fact that the first "improvement" started by the fact that the technical aspect of data collection is seen as a strategic problem. What this assumption can take advantage are the "tools" assembled or governed under a studied strategy for the data collection and exchange. In other terms, the Codacmos integrated data collection process is closely linked to the eGovernment practice.

29.   The main prerequisites for an automated data collection strategy that ease the administrative burden of reporting obligations on businesses and/or respondents and that are considered in the description of the CODACMOS integration model are: 1) the variables to be reported should be registered in their own internal data system; it is intended that the administrative data systems in the private sector (businesses) and government agencies use and develop computing systems; 2) the government shall never request more the information that is in use (already collected); 3) the enterprises shall never have to report the same information more than once; 4) the government shall offer the most efficient means of reporting; 5) there shall be correlation between the use derived by the government from the requested information and the burden it imposes on the business sector. The integrated scenario developed on that basis satisfy at least three of the following criteria: quality (reliability); timeliness; reduction of costs and burden; improvement of efficiency of data collection and the data protection.

30.   The general scenario of the integration model proposed by CODACMOS as in the following shows the overall outline of data collection and exchange between information

systems at businesses and intermediaries and households/individuals and integrated information systems of shared data collectors. From a logical point of view, it is the relationship between the two sides of data collection and exchange process (the origin and the destination) following a set of agreed requisites (objectives) and taking into account that to a certain extent, the issues related to data collection and exchange are similar in all countries and for most of the kinds of surveys or forms. Although the operational (compute or technical) flow diagrams are necessarily different, they share the same conceptual (common) model.

CODACMOS Overall outline

Respondents   **Shared data collectors**

Information System of Businesses and Intermediaries

Single Request

Integrated Information System

Households and Individuals

1 ANSWER
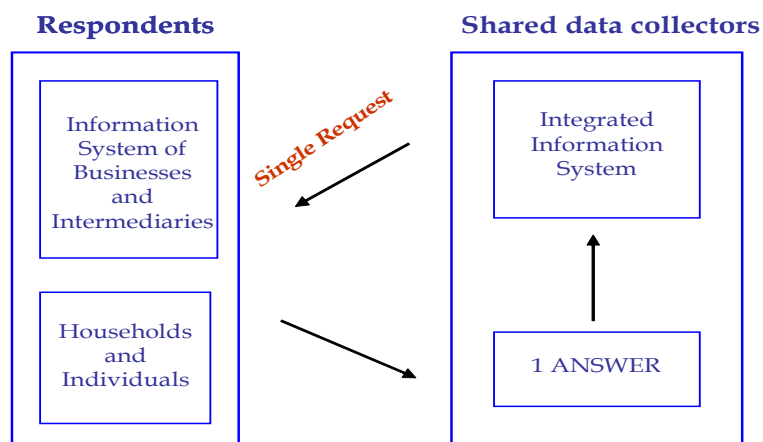
31.   In line with the assumptions as above, below it is shown the data and metadata diagram for a given shared data collector (NSI, National P.A, Social Security, etc.).

32.   At operational level, CODACMOS has considered that the organisational and technological aspects should be associated to the Central Body and Operators in the organisational level that in this scenario are presented in two different locations: central body that contains the organisational aspect of four functions already defined and the management of single request and 1 answer. The software integrated data capture manages the XML files before and after the data collection process. The "Central Data Collection Body" database represents the result of the technological process that is performed when various data collectors (including the NSI) have agreed to formulate their requirements for data (and the relative answer) on single request basis towards the respondents (a data base is designed in order to store both the data and ontology metadata).

Data and Metadata Diagram for a data collector: Conceptual Level
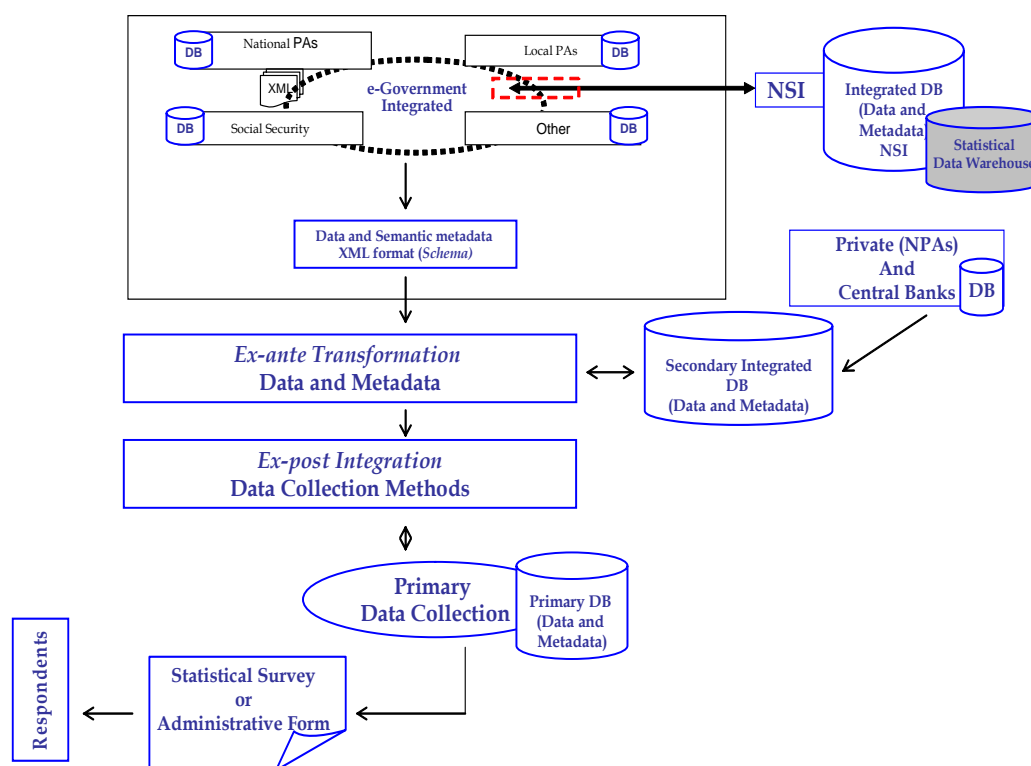
INTEGRATION OF DATA FROM DIFFERENT SOURCES (PRIMARY AND SECONDARY)

33.   The integration of data from different sources is the process which put together statistical and non-statistical data collected and performs burdens relationship among them. The use of administrative data for statistical purposes and the integration of non-statistical data with statistical data doesn't mean always the same thing. The N(data providers) to 1(data collector) CODACMOS approach of integration describes exactly how to integrate the data of secondary sources with the collection of data not available elsewhere (primary) under a given framework and respecting agreed definitions, prerequisites and conditions.

34.   But the integration of data introduces a lot of difficulties in the homogeneity of the information, in the oneness of the common data, in the classification of the registers and archives. Such process is a slow and difficult process for the difficulties that find in the union of data, created for different purposes and for different finalities.

35.   **From the use to the integration of administrative data with statistical data**: Statistics are derived both from statistical surveys and as by-products from administrative data. Much progress has been made in advanced statistical systems in exploiting the latter for statistical purposes. It is a vital area for progress. In particular, computerized data systems must be brought more into play to serve statistical purposes. They are comprehensive,

CES/2005/35
page 14

detailed and a much more economic source of statistics than special surveys. They are also potentially, easy of access. It's obvious that administrative data are a potential gold mine for statistical information, which it might be impossible and certainly less efficient, to collect by special surveys. But, though this may be widely recognised, it's usually hard in practise to capitalize on administrative data for statistical by-products. Administrators often give low priority to the statistical uses of their data, or do not realize how best to exploit them in this way.
Classifications, definitions and coverage may differ from those needed for statistical purposes; and there may be confidentiality inhibitions.

36.   We see more in detail that some problems have to be faced in relation to the integration of non-statistical data with the statistical data:
*   It faces the daunting problem of acquiring the information, it needs to look at current conditions and past trends. Fortunately, there are many sources of existing data that can use for a variety of purposes. Local and state governments collect a lot of data, although much of this may not normally be made available to the general public. There are a number of private sources of data that can be useful. There are several companies, for instance, that sell electronic business directories listing the names, addresses, and types of business establishments.
*   Secondary data have a number of advantages over primary data. First, secondary data is usually less costly to obtain than data you collect yourself. Someone else has gone to the trouble and expense of collecting the data already. This often includes the process of "cleaning" the data, which is, making sure that they contain no errors. Second, using secondary data from a particular source may help create a demand for that data and thereby maintain the data source for future use. Finally, using secondary data can help build relationships between the data provider and the community— relationships that will benefit both groups.
*   With secondary data, it has no control over how the data were originally collected. The types of information obtained, the population sample from which it was obtained, the way that questions were asked, all of this was beyond your control.
*   But there can also be disadvantages to secondary data. Because it has no control over how the information was collected, the data may not cover exactly the population or topics that you want. Data that are collected for administrative purposes, for instance, may not be entirely suitable for evaluating programs. The data also may not be very current or may contain errors. If the data are based on a one-time survey, then it may not be possible to get access to comparable data in the future. Finally, there may be restrictions placed on the use of the information by the collecting entity that make it difficult to integrate the data in practice.

37.   **Collecting data not available elsewhere (primary source):** The integration of existing data is one strategy that users can follow to build up a comprehensive information system. But, despite the wealth of information to be mined from existing sources, such sources will only go so far. It will undoubtedly wish to investigate new areas that are not covered by any pre-existing data sources. In this case, it must consider methods by which it can collect data that cannot be obtained from secondary sources. Unlike secondary data, with primary data it has almost complete control over how the information is collected. It specifies the overall goal of the data collection effort, what questions are asked, who is included in the

sample, and so forth.

38.   The surveys are one of the most common methods of primary data collection. Surveys usually consist of a very structured series of questions that are asked of all participants, and the responses given may be restricted to a set of discrete choices ("very important" , "somewhat important," etc.) or very brief statements.

39.   Primary data have a number of advantages over secondary data. Original data can be tailored exactly to the programs or needs. As a result, original data can provide the most critical indicators by conforming more precisely to the chosen objectives. Furthermore, it is more likely to "buy in" and support a program with original data that irrefutably characterize their unique situation. With sufficient resources, primary data can be made as precise as needed and can be collected as often as needed.[2]

40.   Apart from metadata, several key issues when considering whether a particular source of existing data is appropriate for the integration should be matter of further work:

- The first issue is that of *ownership* of the data. Data from some sources, such as commercial vendors or local government agencies, may be private or proprietary. In this case, you might not be able to obtain all of the data that the provider collects (for instance, certain pieces of identifying information may be purged from welfare case records), or there may be a licensing fee to get access to the data. Furthermore, the provider may put certain restrictions on how the data can be used and whether they may be shared with others. Some private vendors, for example, allow only the licensing organization to access the data—the information cannot be provided or reported to others.

- Related to the issue of ownership is the question of *confidentiality* of the data. Some information collected by government agencies for administrative purposes may be sensitive and, to protect the privacy of their clients, not releasable to the public. In this case, there are two options. The provider can simply omit the sensitive information (names, addresses, social security numbers) from the data before releasing them to others; alternatively, the provider can summarize the data so that it is not possible to identify individual cases. For example, welfare caseloads could be summarized at the neighbourhood level. These approaches solve the confidentiality issue, but can make the data less useful. For instance, it might want to use social security numbers to match welfare case records with other types of social service records. If this information was removed by the data provider it will not be possible to do this.

- A second option that gets around this problem is for the community group to enter into a *confidentiality agreement* with the provider. In this case, the community group would get access to the complete, uncensored version of the provider's data, but certain restrictions would be placed on their use for integration. The agreement might specify that the data can only be reported in a form where individual cases cannot be identified. It might also require the community group to keep the data in a secure location (such as on a password-protected computer).

- Another issue to consider is the *timeliness* and *frequency* of the data. Are the data based on current information? Or are they out of date? Are new versions of the data collected on a regular basis? Of course, one always wants to have up-to-date data. But this may not always be possible. So, to compare populations across cities, it may be able to use very recent data, but neighbourhood comparisons may have to rely on

older information.
- This leads to the issue of *geography*. Data from some sources may be available for small levels of geography (blocks, Census tracts), while other data may only be provided for larger areas (cities, counties, states). Survey data are especially difficult to get for small areas, because to be able to produce accurate estimates it has to have many observations in the survey sample.
- A final issue regarding secondary data is the *format* in which the data are provided. Do the data come in a file format that can be readily read into a PC program (such as Excel or Access)? Or is it an ASCII file that must be converted? Is there documentation clearly identifying each of the data fields? Do the data come on diskettes, CD-ROM, or computer tape? How large are the data files? Do they require a lot of processing before they can be integrated (such as summarizing or combining)? All of these questions are important to answer up front so that it do not invest a lot of time and effort obtaining data that it do not have the technical resources or expertise to integrate.

INTEGRATION OF PRIMARY AND SECONDARY SYSTEMS FOR A NATIONAL STATISTICAL INSTITUTE

41.   The essence of a good statistical information system lies in the coordination and integration of data and metadata and relative systems of primary and secondary. In the past, the task of the statistical office was seen as producing facts on each of a wide range of subjects, with little attention to connections and integration between them. The starting point was the national accounts that changed all that. One of the remarkable developments in government statistics has been the building up of inter-connected statistical systems, above all in the economic accounts, but also increasingly in social and demographic statistics where was integrated primary and secondary systems.

42.   It's no longer necessary to demonstrate why this is a major advance. Facts again disproportionately in usefulness the more they are related to other facts, and it is the miracle of the economic accounts that they reflect in sophisticated detail the complexity of the real world. Moreover, social and economic analysis and policy itself has become increasingly concerned with inter-connections whether between the "ripple effects" of any particular policy or the inter-related implications of different policies. Numerical models have been developed to enable such analysis to be made in a precise manner. The official statistician must concern himself with systems, linkages and models, and, generally, with all aspect of the integration data.

43.   This is not novel for national statisticians which have participated in developments in international gatherings. Even so, the NSI itself does not yet place the integration and co-ordination of statistics, linkages, models and, generally, the building up of statistical systems, at the centre of its concerns. In most advanced statistical systems, all specific developments and priorities are now viewed as part of an overall integrated system. This is also the approach of NSI and priority should be given to perfecting the tools and solutions of integration: the accounts themselves, classifications, standards, registers, etc, in reference to primary and secondary systems. The reality in the European Countries is very different (differently of other countries, in the Nordic countries the statistical information systems integrate mostly secondary data systems). However, the building of statistical systems and the

integration of data generally, is matter of fact in the work of NSI. The construction and maintenance of registers, especially of business establishments, is of outstanding importance.

44.   In that framework, the integration of primary and secondary systems requests the creation of the kind of unit with the sole function of planning all aspects of co-ordination and integration within NSI, (which can be extended to other NSI and other central agencies; and between and within regions and other geographical grouping). In this process the role of the NSI becomes very important.

45.   From the point of view of a NSI, the definition of integration system of primary and secondary data and metadata could be mainly meant as the obtainment of a set of requisites and relative solutions that satisfy as much as possible the requirements of diversified users that are more and more demanding to a better quality, timeliness and completeness of statistical data and products, that cannot be guaranteed by means of only one information system source in the context of the actual ICT developments. This, obviously, represent the most flexible definition that sufficiently describes the actual necessities. This definition fits the described given to the data collection strategy reported in other Codacmos products (deliverable 3.2 and other).

46.   This framework definition emphasizes the importance of work on methodology within a statistical office, or, in general, in the office responsible of integration process. Data quality is crucial and requires strong methodological back-up. Not solely mathematical statistical work, but, i.e. all aspects of survey and sampling design; quality control, the methodology of sampling frame; data linkages techniques; the developments of computer systems; index construction; data exchange, capture and editing; estimation procedures; seasonal adjustments, model buildings, for economic, social and demographic applications; and so forth. It can't emphasize enough the role of methodological work in designing and building statistical systems.[3]

47.   A concrete example of integration of primary and secondary system is the Statistical Information System on Enterprises and Institutions (SISSIEI) of Istat.

INTEGRATED SOLUTION OF PRIMARY AND SECONDARY DATA COLLECTION

48.   The main questions that CODACMOS working groups have identified on this issue could be addressed as in the following:
•   Are the Primary and Secondary data collection basically processes that could be seen as independent objects with some known relation to each other?
•   Are there some integrated solutions?
•   On what level integrated solutions?
•   These questions yielded the following answers.

49.   More or less sophisticated or universal solutions always exist and the key issues for them are legislation, data and metadata standardization and inter-institutional data sharing. Until it adopts solutions that are more sophisticated, the best practice approach could be applied.
50.   They should be considered and hence analysed as independent objects as the data

content often differ (e.g. primary data are usually collected as raw, non aggregated data). On the other hand secondary data has often embedded in itself the specifics of its data collectors as for aggregation, summarisation, system tools and technology used etc.
The integrated solution has at least three aspects:
i)   *Integration at data level* - this should be considered as a primary goal in primary and secondary data collectors. Tools for supporting the data integration are standards both for data and metadata;
ii)   *Integration at process level* - is supported by data and metadata standards and technology standards. Important are standards on classifications and code lists, less important are standards of technology used;
iii)   *Integration at state level* - is supported by the standards mentioned above and complemented by legislation and state level regulations. Mutual agreement among the institutions at state administration level on data mining or data transfer could substitute the legislation in case of lack of it. Partly data protection and other regulation may also prohibit non-legislated data sharing.

51.   An effective integration is a combination of all three aspects. In the following there are described some considerations related on the levels of integrated solutions.

52.   **For a NSI, the integrated solution at data level** or the "intra muros" integration is intended the process that is performed within the organization. This type of process, if completed alone, it cannot bring to an elevated improvement of the quality of the data. It should be strongly connected with the "extra muros" integration process.

53.   The process is subdivided into 2 phases: "ex ante" and "ex post" integration. The ex ante integration includes the implementation of:
*   Integrated registers; they are necessary to load all information which become from primary and, eventually, secondary data capture process. These registers must contain all information to classify the micro and macro data.[4]
*   Metadata model. The metadata model is necessary to load all variables which are captured by primary and secondary data collection and estimated variables.
*   The ex post integration includes the implementation of:
*   The load process of all information deriving from data providers;
*   The integration of data coming from the different sources.

54.   Two examples that CODACMOS have considered are: SISSIEI system that integrates all information deriving from surveys, administrative sources and registers and the consumer price system (SISPRE).

55.   As it is known, a statistical system aims to provide knowledge about the values of statistical characteristics, i.e. characteristics which give a quantified, descriptive summary of the relationships in a group. A statistical system consists of, in an early stage, the collection of information about individual objects, which in a later stage, is processed into statistics. In an integrated statistical system it's difficult to delimit the single surveys associated with a given documentation, quality declaration, or some other description. Analogous difficulties in "delineation in time" can also occur, especially with continuing periodic surveys.
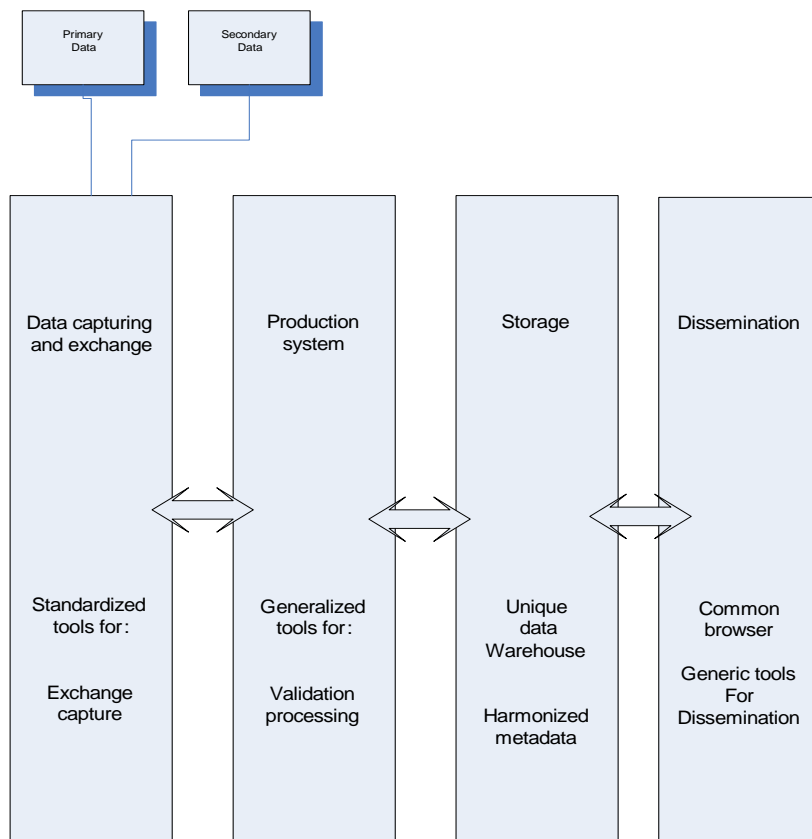56.   The development of a statistical system is a part of a larger strategy aiming at the

coordination of the production of statistical data. In particular, the integrated statistical information system is intended as a super-system which gives the possibility to integrate all available statistical information relative to a single statistical survey, administrative data, etc. The system should be designed to permit the comparison of data referred to the same variable from different surveys (monthly, annual, etc.) for each statistical unit, giving the possibility to analyze also the evolution of the unit in terms of mergers and acquisitions, to check data quality, to prepare products for dissemination, to conduct micro econometric analysis, etc. within the same information system, with a clear reduction in costs and a large improvement in the efficiency of the statistical production. In addition to, the statistical system should give internal users the possibility to access generalized tools to conduct surveys, like software for sample analysis, to manage questionnaires and contacts with enterprises by post, fax or e-mail, to check data, etc.

57.   The integrated solution at data level considers the tools, standardization and metadata. It comprises four components:
- data capturing and exchange tools;
- production tools;
- storage system by means generalized tools, harmonized metadata system, and unique data clearing house;
- data dissemination tools.

58.   The figure overleaf shows schematic view of the architecture of the integrated solution at data level. This solution can be implemented in the conditions that provide enough possibilities to integrate data and metadata between primary and secondary sources. This is a flexible system that can be run in centralized and decentralized systems due to new developments in information technology offer advanced tools for managing statistical systems that employ integrated solutions of primary and secondary. This architecture provides the possibility to NSI to supply data based on XML and associated technology standards, by posting data on their web sites using a standard XML based format (data supply to international organizations in particular).

```
┌──────────┐   ┌──────────┐
│ Primary  │   │Secondary │
│  Data    │   │  Data    │
└──────────┘   └──────────┘
```

Data capturing and exchange	Production system	Storage	Dissemination
Standardized tools for:	Generalized tools for:	Unique data Warehouse	Common browser
Exchange capture	Validation processing	Harmonized metadata	Generic tools For Dissemination

59.   **The integrated solution at state level or the extra muros integration** is intended the process that is performed in common with other organizations and institutes. It guarantees a better improvement on data quality. The process is subdivided into 2 phases: "ex ante" and "ex post" integration. The ex ante integration comprises all phases of intra muros process, but these are conducted with external Institutes, so the registers, the metadata model, the nomenclature have to be completed in common agreement with all partners. This task is more complicated by the organizational point of view, with an increase of difficulties from the architecture point of view, with more complication since it is necessary to exchange data in a secure mode. There are some steps to put in common vision before the process of exchange starts. In the development of the extra muros process is necessary to make reference to a unique statistical process, much more complicated than an intra muros process. In fact, it is necessary to consider the structure of the files to exchange, the channels in use and the security infrastructure context. Furthermore, an optimized process needs that not all the available information is exchanged, but only the one that is essential to the productive process. In that aspect the request for information to be exchanged should be satisfied in real time, using the infrastructure web and using the necessary resources to the security, as the PKI, like for instance the structure SSL, and so on.

60.   The examples taken into consideration from CODACMOS project on that aspect are: the TYVI solution (Finland) and the software developed by Codacmos partners for data exchange in secure condition (ISTAT-INPS).

61.   Some important issues that are related to the integrated solution at state level could be

listed as follows:

62.   **Standard interfaces:** The use of more extensive systematic use of standard interfaces lead to the following improvement:
- a drastic decrease in the complexity of data exchange between statistical information system and their environments as well as between the subsystems of the individual statistical information systems themselves;
- a drastic increase in the variability and flexibility in the behavior of the statistical information systems.

63.   The use of standards permits that every system will be able to communicate with all other systems, including systems that are do not yet exist but will be introduced later. This situation permits both fewer complexes to develop, maintain, and operate, and more flexible vis-à-vis growth and other changes in the system environment. The standardization may distinguish between external, inter-system interfaces, and internal, intra-system interfaces.

64.   **A database oriented to internal and external interfaces:** Today the relational data model and the SQL standard for data interchange between application software and the data base management system are obvious choices for internal interfaces. There are no the same standards for any period, but this is not a problem.

65.   **Standard components:** the NSIs were among the first companies and organizations to make systematic use of standard components in the development of information system applications. With the advent of inexpensive personal computer technology and software, the boundary between user programming and professional programming has become blurred; in statistical offices as well as in the data processing community at large. The paradigm shift is likely to imply an even greater future for such things as inexpensive, generalized software, available "off-the-self", "tool-boxes" containing generalized standard components, and Rapid Application Development (RAD) methods and tools.

66.   **The harmonized metadata:** One of the major obstacles of data exchange remains the inadequacy of available metadata, that is, the absence or inadequacy of systematic descriptions of statistical data and the process behind them. For the data exchange process it has need metadata for three major purpose: searching for potentially relevant and useful statistical data, evaluating the adequacy of available data and the cost / benefit of using them, and, finally, retrieving, interpreting, and analyzing statistical / administrative data. The NSI needs to integrate statistical and administrative data from several sources, and for the present there is no organizational unit within a National Statistical Institute that has the necessary overview. From user's point of view, it expect the metadata needed to be organized and disseminated in such a ways that himself can look for relevant data on the basis of widely available, computerized metadata.

67.   For extra muros data exchange process it is requested to determine if the data are really adequate for the intended purpose. This means that it should be evaluated the quality of the data, and to consider whether it is really the case to put money to retrieve, interpret, and to analyze the data. Interpretation and analysis requires the same kind of metadata as the ones needed to make the preliminary assessment on the quality of data. However, it may be necessary to obtain deeper and more precise information about how the data were collected

and processed, before they are transformed in available statistics.

68.   The metadata describing a statistical survey and its data outputs are a combination of formalized metadata and free-text metadata like verbal description of variables and process. The information system for handling statistical metadata may require different types of software components to be integrated, i.e., relational database management software for managing and searching large amounts of text data.

69.   For the description component of metadata the semantic form is not enough to exchange the data in the case of extra muros integration, in this case it becomes necessary to develop an ontological description of the same meaning.

70.   Another task to be developed is the dynamic metadata concept. Dynamic metadata represent the concept that in data exchange process for the organization which send the information, one variable may represent a micro level data; for the organization which receives the information the same variable is managed as macro level data. The definition of metadata has to be changed, so isn't simple to decide if this variable, at general level, has to maintain the micro or macro content.

71.   **Confidentiality:** Statistical data can only be made available to the users within the limitations of certain confidentiality restrictions. The most fundamental purpose of these restrictions is to preserve the data provider's confidence in the statistics producer's willingness and ability to ensure that data submitted to a statistical producer will be used for statistical purposes only. Among other things the statistics producer must be able to ensure that statistical outputs will not, thanks to the input submitted, directly or indirectly, enable a statistics user to associate sensitive information with the data provider or anyone whom the data provider would like to protect. Statistical confidentiality can only be ensured by a combination of technical and legislative actions.

REFERENCES

CODACMOS project, Deliverable 3.1.1 - "Report on the 1st Technical Meeting of the Working Group on Primary Data Collection Athens, 29-30 May 2003".

CODACMOS project, Deliverable 4.1.1 - "Report on the 1st Technical Meeting of the Working Group on Secondary Data Collection Athens, 29-30 May 2003" .

CODACMOS project, Deliverable 3.1.2 – "Advanced results of the Working Group on Primary Data Collection and report on the 2nd Technical Meeting, Rome, 21-23 January 2004".

CODACMOS project, Deliverable 4.1.2 - "Advanced results of the Working Group on Secondary Data Collection and report on the 2nd Technical Meeting, Rome, 21-23 January 2004".

CODACMOS project, Deliverable 5.1.1 - "Report on the 1st Technical Meeting of the Working Group 5 on the Integration of Primary and Secondary Data Collection Helsinki, 3 September 2003".

CODACMOS project, Deliverable 5.1.2 - Report on the 2nd Technical Meeting of the Working Group on Integration of Primary and Secondary Data Collection, Edinburgh, 28-29 April 2004.

CODACMOS project, Deliverable 7.1 - Report on the key list of issues/problems for further European research supporting efficient data collection for statistics and related metadata.

CODACMOS project, Deliverable 3.2 "Report on definition and the model on primary data collection and intermediaries".

CODACMOS project, Deliverable 4.2 "Report on the common model on secondary data collection and guidelines how secondary data will be used".

CODACMOS project, Deliverable 5.2 " Report on guidelines, recommendations and development of integrated solutions for primary and secondary data collection.

CODACMOS project, Deliverable 6.2 "Report on "common" core model in data collection and on the definition of level of standardisation".

---

[1] ISTAT actually doesn't have such a coordination unit.

[2] But, of course, primary data also have their disadvantages. The major disadvantage is the cost and effort that are required to collect primary data. This is especially true for neighbourhood-level data, where a large number of observations are needed to obtain sufficiently precise estimates for small areas. Collecting new data can absorb valuable resources that might otherwise be devoted to other efforts, and, once collected; primary data create a recurring demand for more data. Finally, certain primary data collection methods require technical expertise or resources that are not readily available in some communities.

As it is known, there are several key problems that must be addressed when considering whether a particular method of collecting data is appropriate for primary use:

- The first issue, is that of *cost.* Collecting primary data can sometimes be very expensive. There is a great deal of time involved in doing, for example, door-to-door interviews. If it need to pay the interviewers, then the cost will be greater the more "doors" it need to visit. To do a telephone survey, it may need to pay an organization with a computer-assisted telephone interview (CATI) system to conduct the survey for you. Even focus groups can be expensive, if they involve mailing out large numbers of invitations to prospective participants and hiring professional facilitators.
- A second issue is that of the *technical difficulty* in doing primary data collection. Most data collection methods require specialized knowledge and skills to carry them out properly. For example, in surveys it must decide what population it has trying to collect data on and then choose a proper sample that will give useful information on that population. Knowing how to construct a proper sample and interview the right number of people requires some expertise in statistics. Furthermore, it must design survey questions that will accurately obtain information on the issues in which you are interested and carefully train interviewers so that they administer the survey properly.
- One must also be concerned with the unit of *geography* when collecting primary data. If you want data for an entire city, then you must cover more area than if you need only data for a single neighbourhood. But, if you want data both for the entire city *and* for individual neighbourhoods, then you will need to structure your sampling and data collection differently.

Finally, as with secondary data, there can be issues of *confidentiality* associated with primary data. If you are asking people to reveal personal information or to respond to sensitive questions, it will probably need to provide assurance that this information will not be revealed publicly in a way that would allow someone to associate specific answers with a particular person.

[3] In general, the NSIs devote a fair amount of attention to methodology. In our opinion, it should be a separate

department devoted to research and development, but it cannot judge how effectively it works or how well its findings are integrated into the regular work of NSI.

From the organizational aspect, a methodology group, covering economic statistics and social statistics and responsible for ensuring co-ordination and overall data quality, and a strong background for methodology is the main priority settled that could be the tool to reach the desired level of integration of primary and secondary systems.

[4] In Italy for example, every year it is build a register which contain all active enterprises. This list derives from the process of integration of registers supplied by the Ministry of Finance, Chambers of Commerce, Social Security, National electricity provider, telephone yearbook, and other registers. The probability to have, at final stage, a complete list of enterprises becomes high. Another example of integrated registers is the code list of economic activity (ATECO). This code list is constructed by collaboration of some Institutes; Istat prepares the first draft which is controlled by specialists which belongs to other institute, for example Ministry of Finance, Chambers of Commerce. The final list comprises all possible economic activities.

* * * * *