



**CAENTI**  
Coordination Action of the European Network of Territorial Intelligence  
A project funded under FP6 of the E.U.  
<http://www.territorial-intelligence.eu>



# **CAENTI**

## Coordination Action of the European Network of Territorial Intelligence

Specifications for the “PRAGMA” data collection and  
quantitative treatment software.

March 2006, 1<sup>st</sup> – December 2006, 31<sup>st</sup>

**Deliverable N° 54**

Jean-Jacques GIRARDOT  
Work package leader  
Université de Franche-Comté

**December 2006, 31<sup>st</sup>**

## Summary

<b>Introduction</b>	<b>4</b>
<b>1. Basic principles of PRAGMA design</b>	<b>8</b>
<b>1.1. An intuitive use</b>	<b>8</b>
<b>1.2. A collective tool</b>	<b>9</b>
<b>1.3. The questionnaire as a structuring data processing document.</b>	<b>11</b>
1.3.1. Empirical constraints linked to the treatment.	12
1.3.2. PRAGMA formal principles	13
<b>2. Questionnaire Specifications</b>	<b>15</b>
<b>2.1. Access rights</b>	<b>16</b>
2.1.1. Consultation profile	16
2.1.2. Key-in profile	16
2.1.3. Master profile	17
<b>2.2. Questions specifications</b>	<b>18</b>
2.2.1. Question row	18
2.2.2. Question formulation	18
2.2.3. Question Code	18
2.2.4. Question form	18
2.2.5. Modality without answer	19
2.2.6. Modality without object	19
2.2.7. Question with unique or multiple answer	20
<b>2.3. Specifications of the modalities</b>	<b>20</b>
2.3.1. Modality row	20
2.3.2. Modality mnemonic code	20
2.3.3. Modality Type	21
2.3.4. Modality frequency	21
<b>3. Specifications concerning data</b>	<b>22</b>
<b>3.2. Code</b>	<b>26</b>
3.2.1. Digital code	27
3.2.2. Mnemonic codes	27
3.2.3. Boolean codes	28
<b>3.3. Studying and coding values</b>	<b>28</b>
3.3.1. Measures division into classes	29
3.3.2. Gathering texts into categories	30
<b>4. Specifications of controls for robustness and data quality.</b>	<b>31</b>
<b>4.1. Automatic controls</b>	<b>31</b>
4.1.1. Answers absence and question without object	31
4.1.2. Unique answer and multiple answers	32
<b>4.2. Coding and recoding functions</b>	<b>33</b>
4.2.1. Control of the missing answers	35
4.2.2. Coding values	37
4.2.3. Recoding	38
4.2.4. Selecting characters for qualitative data analysis	40
<b>5. Other data analysys and data processing specifications</b>	<b>41</b>
<b>5.1 Flat sorting</b>	<b>42</b>
<b>5.2. Global quantitative balance</b>	<b>43</b>
<b>5.3. Cross sorting</b>	<b>44</b>
5.3.2. Selection	44

<b>5.4. Data sheet</b>	<b>45</b>
<b>5.5. Synthesis question</b>	<b>45</b>
<b><i>Conclusion: Prospects</i></b>	<b>49</b>
<b><i>Bibliography</i></b>	<b>52</b>

## ***INTRODUCTION***

This report aims at defining the technical specifications of the PRAGMA software so as to direct the evolutions of the data-processing developments of the “CATALYSE toolkit”. The latter aims at offering selections of information, which are homogenized at the European level, according to the CAENTI actors experiment and to the existing European standards, which allow confronting at the territory scale:

- The people’s needs that are gathered by means of a diagnosis and evaluation guide that presents under the form of an individual questionnaire;
- The available services to meet these needs, which are inventoried and updated in a services repertory that is a database that is published on line;
- Territorial indicators that are gathered and updated in a territorial information system, which can possibly be published on line.

The WP6C and WP6G coordination groups drafted contents specifications of the guide, the repertory and the contextual indicators. The deliverables “European contents specification for a CATALYSE guide for diagnosis and evaluation” (deliverable n° 51), “Guidance notes for the use of CATALYSE information and tools” (deliverable n° 56) and “List of territorial indicators available on internet for comparison with CATALYSE Guide Data” (deliverable 53) already offer elaborated versions of the guide and of the repertory and a selection of territorial indicators. With the papers “Specifications of the contents of the European guide of diagnosis and evaluation” [SANCHEZ, GIRARDOT, 2006] and “Activities and prospects of research activities concerning tools of territorial intelligence for sustainable development actors” [GIRARDOT, 2006] they give the useful information about CATALYSE method, about the guide of diagnosis and evaluation, and about the drafting of the European guide in the framework of the CAENTI.

The project of « CATALYSE toolkit » also aims at improving the tools accessibility by keeping simplifying their use and by automating the treatments. That is why the contents specifications also concern the definition of the treatment protocols (see deliverable n° 56). The latter ones are a modeling of processing steps that allows automating some treatment phases. They also allow a better integration of the tools, what is the issue of the deliverable “Specifications for the PRAGMA integration with the software of qualitative data analysis ANACONDA and NUAGE” (deliverable n° 55).

PRAGMA is the most popular tool of the CATALYSE method, because it allows making the key-in of the data that were gathered with the diagnosis and evaluation guide, then the quantitative treatments, which are called sorting by the statisticians, which results are published in the form of frequency tables, and then of histograms or maps. It also allows structuring the data table, standard format that is used by the data analysis software, in particular ANACONDA.

Usually, PRAGMA is used within development partnerships to digitalize the guide in a questionnaire form after the partners have defined it. Then, it is divided between the latest so as every one key-in, directly or not, the answers that were gathered during the interviews. The data are periodically regrouped for the global treatments. It allows calculating cross-tables, selecting data sub-sets from which it is possible to make the same treatments or to prepare the qualitative treatments.

The initial project of the “CATALYSE toolkit“ is to dissociate the “gathering” part from the “treatment one”. It allows lighting the “gathering” part of the treatment functions. The gathering is the task that is divided between the partners and that is executed by little-skilled users, whereas the treatment requires data processing and data analysis skills. To isolate PRAGMA “gathering” part, we suppress the functions that are linked to the treatment, what limits the risks of handling errors and of data corruption.

This objective was reached insofar as a PRAGMA version that is limited to the gathering was made since May 2006 for the key-in of the “migrants” guide. The latter is a homogenized synthesis at the Spanish scale from the guides that are used by four migrations observatories. This synthesis was a step of the homogenization process of the European guide within the CAENTI. The ACCEM observatories development allowed experimenting the guide in two new observatories with a PRAGMA “gathering” version which specifications were defined by the WP6P coordination group, which is in charge of the technical specifications.

According to the opportunity of jointly experimenting the « migrants » guide in four former migrations observatories (GIJON, OVIEDO, SIGUËNZA and GUADALAJARA) and in three new ones (SEVILLE, LEON and GIRONA), a PRAGMA “gathering” prototype was made in this framework and it is being experimented in the seven observatories that are animated by ACCEM in Spain [FERNANDEZ, J., MAHIA, J., & al., 2006].

This possibility to test out a diagnosis and assessment guide in the same language gives an opportunity that we initially did not consider because it results from the recent dispersion of migration observatories in Spain at the initiative of the partner ACCEM. The development of multilingual versions is essential for the technical translation of specifications in terms of tools, then for the experimentation of contents and advised tools, because the hands-on actors rarely practice several languages. The development of a multilingual version of PRAGMA, which constitutes an inherent specification, presumes a new software conception. Testing out a diagnostics and assessment guide and synchronized gathering and analysis tools, allow us to progress quickly about the validation of guide contents and the establishing of processing protocol, about the technical assessment of tools, and more globally about the uses observation.

These experimentations dynamized the demand as regards the PRAGMA software within CAENTI, on the one hand to make a multi-platform version and on the other hand to start developing an online version that was planned for 2007. They are also important, as for the software technical developments, as to test its using and its use, even the definitions and the suggested treatment protocols accessibility or to delineate the necessary trainings and accompaniment.

Some of the specifications that this report presents gather and synthesize the PRAGMA experience since its creation in 1991. They constitute principles to follow in order to remain in the spirit of a tool pre-eminently oriented toward conviviality and accessibility. The others are recommendations about the development of PRAGMA. They remain voluntarily general, without precise technical translation, for two reasons. Some of these recommendations have already been experimenting as we have just explained. It is important to test the technical aspects but also the contents and above all the uses to validate these specifications. The others refer to the integration with the other analysis software within an information system and the on lining of PRAGMA. Thus, they refer to very different technical developments. They should remain enough general to be useful to the development of very different versions at the technical level, but of which use should remain the same, whatever is the used technical tool.

Firstly, we will present what we will call the basic principles of PRAGMA design. They compose a set of fundamental orientations that since its design made it a friendly, economic and easy-to-use, an intuitive tool that was designed for a collective use that was

based on a specific conception of the questionnaire in the prospect of its statistic exploitation. The future developments will have to respect them to remain in the PRAGMA philosophy.

Then, we will present the specifications about the questionnaire: rights, questions and modalities.

In a third part, we will evoke the specifications that are linked to the data coding. They are a fundamental aspect of the precise and strict formalization that allows implementing the quantitative sorting and integrating the treatments, in particular the quantitative, qualitative and spatial ones.

Then, we will present the control procedures that guarantee PRAGMA robustness. They prevent the manipulation errors and guarantee the data quality. Some of these controls are automatic. Then, PRAGMA has functions series that allow making evolve the questions towards a form that is appropriated to the wanted treatments. They integrate in a data qualification chain so as to allow simplifying and accelerating the quantitative sorting, and then preparing the qualitative treatments.

Finally we will present the processing specifications. It is on this level that we will approach the specifications concerning the future developments of PRAGMA.

## ***1. BASIC PRINCIPLES OF PRAGMA DESIGN***

PRAGMA is a software of questionnaires surveys analysis, which is directed towards the quantitative processing. It was initially developed in 1991 [GIRARDOT, 1991] to make the coding operations that are necessary:

1. To make simple and cross quantitative sorting;
2. Then, to make the data table in the prospect of the data qualitative analysis that will be made by ANACONDA, software for data analysis.

Besides, PRAGMA design answered several needs that it was possible to jointly meet thanks to the fabulous intuitive and interactive potential of the HyperCard software. Since 1984, the personal computer “Macintosh” by Apple popularized a graphic interface that notably simplified the computers use. With HyperCard, it became possible to develop intuitive applications that are closed to the working practices.

PRAGMA was designed for a collective use, to mutualize and share the data upstream, and to collectively analyze the results of the data analysis processing downstream.

So as to favor the access of an increasing number of users to the information quantitative and qualitative analysis techniques, but also to face the demand of databases statistical exploitation, PRAGMA integrates control procedures of data quality. They are based on a questionnaire formalization as a document of data structuring in preparation for the statistical processing. Thereby PRAGMA pays a detailed attention to the data coding that depends on the form of the questions, independently of their meaning, and especially of the form of the answers.

### ***1.1. An intuitive use***

PRAGMA displays each question on the screen with a similar presentation to the one appearing on the paper questionnaire. We key-in an answer by clicking on the modality that corresponds to it in the modalities list that is on the screen. In the case of an open question, we key its value in, as we usually do with a word processor. At any moment, we visualize the list of answers made by a person, by clicking on its code. It is possible to go back to a question to consult the modalities frequency or to establish an index of the values that allow consulting the persons who made the same answer. To code a question, it is necessary to select values that are close, and then to constitute a new modality with some clicks. In the same way, we gather several modalities in a modality. The answers are summed in real time

at each question level: by consulting, or possibly printing, the screen of a question to have a corresponding counting table, for all the individuals or for a selection. The simple or cross tables are registered in files that have a text format. A spreadsheet, such as Excel, allows reading them, completing the calculations and drawing varied kinds of histograms.

PRAGMA affirmed itself as simple-to-use software. The keying-in became an operation that can be made by any person who knows how to use a word processor. The researchers and the professionals could make the coding, recoding, selection, statistic sorting operations, and then complete the exploitation and communication of the results thanks to usual office automation tools.

### ***1.2. A collective tool***

With the diffusion of the micro processing, the data-processing users circle notably widened and the size of the databases strongly increased. At the beginning of the 1980's, important data sets were collected "on the ground" thanks to surveys or observations. The laboratories of Geography, Archaeology and Chrono-ecology were particularly specialized in the systematic gathering of data sets about several thousands of persons or observations. The gathering needed the coordination of several teams. The new users of data processing were neither data processing specialists nor experienced users of software. They worked in varied data-processing environments. The shared keying-in allowed making voluminous bases that combine skills of several disciplines or of several activity sectors. Results analysis and interpretation have also become multi-discipline and multi-sector cooperation. The HyperCard, and then the ToolBook, software, had the advantage to gather in a same file all the software functions and data. It allowed gathering data entries that were made by uncollected users without any complication, very closely to the survey or observation places.

Thus, PRAGMA allowed leading the gathering of data about more than 2155 descriptions of archaeological establishments, so as to cross them with environmental bases corresponding to an area of 15000 km<sup>2</sup> in the Rhone Valley (in a triangle between Lyon, Nice and Montpellier in France), in the framework of the Archaeomedes programme of fight against soils desertification in Europe (1992-2000). Then, it was used to establish the social map of HUELVA (Spain) with multi-sector data about 3852 households; the diagnosis of the Departmental Plan of insertion of Doubs (1994) about 8897 households or the evaluation of 16 "Boutiques Solidarité" of the Abbé Pierre Foundation concerning 5895 homeless in France (1995). These bases were constituted by tens of teams of researchers or actors. They defined

and validated a common questionnaire that integrates several disciplinary themes or several sector approaches. This questionnaire was registered in PRAGMA. A virgin specimen of data was distributed to each team. Each of them gathered and keyed-in their data in a partial file. These files were gathered in a sole PRAGMA basis, that is similar to the partial bases but that includes all the statistical individuals. There were global and partial analyses and possibly a confrontation with other data. The teams that shared their data were associated to the data analysis and interpretation.

The PRAGMA friendliness lays on the reproduction of the usual working environment by the data processing tool. Then, it is based on the integration of PRAGMA in the office automation environment. Thus, it prepares a transition towards new uses the usual working environment cannot provide.

PRAGMA also emphasizes the collective dimension of friendliness that is based on the sharing and cooperation concepts. Presently, we become aware of its implications.

1. It implies the constitution of a community that is united by a common project. This community has no biological, historical or religious dimension; it is simply united by a common objective. The information analysis does not necessarily constitute the project. It is an intermediate step, an instrument, which is at the service of a more general objective.
2. It needs an organization to define the tasks that contribute to the objective execution, to make an inventory of the useful means, to establish a working calendar and to implement a communication. Nevertheless, a decision-making hierarchy does not always structure the community. Communication is not channeled from the top to the bottom. It also works upwards and in a transversal way. The members' rights are equal, what implies a proposal and validation process in the decision-making.
3. Each community member should find its advantage in cooperation without prejudice for the global added value. Thus, beforehand it is imperative to define each ones' right on the shared information. The information contribution and the time that is devoted to it imply a restitution that interests the participants on a collective and personal basis. The participation to the results analysis and interpretation are the complement of the information sharing.

4. The improvement of the information, data and results accessibility is another condition of the community members' involvement.
5. As well as the transfer of scientific and technical skills about the methods, protocols and tools of analysis that are used.

PRAGMA should also provide the possibility to statistically exploit databases. Diffusion of micro processing in the 1980's also contributed to the important development of databases. Many firms and administrations constituted bases that can be used to make statistical analyses. Nevertheless, they were not designed in this prospect. It was necessary to restructure them to respect the constraints of a statistical processing.

### ***1.3. The questionnaire as a structuring data processing document.***

The PRAGMA specifications are based on the qualities the questionnaire should have as a document structuring the treatments, insofar as it formalizes the data for their statistical processing.

In the formal level, the questionnaire is a document that is composed of questions. It is written before making the survey because the pollsters will have to respect the questions number, order and formulation and to survey all the persons in identical conditions, so as the answers variations cannot be attributable to the way of the interrogation set. The situation is different from the interviews one, in which even if we list the topics to be evoked in a guide, the pollster keeps the liberty not to evoke some topics, to modify the order with which they are evoked and to modify the formulation to better lead the interview or to adapt to the surveyed person.

PRAGMA essentially considers the questionnaire as a tool of processing structuring that allows formalizing the data so as to optimize the treatments quickness and quality. It clearly dissociates this treatments structuring downstream function from its communication upstream function. The questionnaire allows a dialogue, which is granted strongly formalized and sometimes distant, between the pollster and the polled in the prospect of gathering the answers of the latter. PRAGMA considers the questionnaire design and drafting as a communication tool are a preliminary and external task.

Thus, the first task PRAGMA offers is the questionnaire recording (or digitalization), what implies it is wrote. The efficiency objective at the treatment level implies the questionnaire design is close to formal principles that should inspire the questionnaire design,

but they can prove to be contradictory with the communication objective. In particular, an open question that allows a free answer is preferable as regards communication with a closed question, which limits the response to the choice between preset items. But only the closed question makes it possible to carry out statistic sorting.

Nevertheless, PRAGMA does not freeze the questions form according to these principles. On the contrary, it considers the question form evolves, from a form that is appropriate to communication into a form that is better adapted to treatments. Thus, for example, we can key-in any question under an open and/or a close form, what allows evolving from the answer free expression towards a series of modalities that are statistically representative that it will be possible to quickly sort and count.

The communication objective is not abandoned because:

- A clear upstream communication is a condition for a good data analysis;
- We will also be concerned downstream by presenting the results under a form that facilitates communication and understanding.

However, in the event of contradiction, the efficiency of the processing remains the priority criterion of PRAGMA.

### **1.3.1. Empirical constraints linked to the treatment.**

These specifications are pragmatic lessons that are drawn from the analysis of voluminous questionnaire surveys and from the restructuring of databases in the prospect of their statistic analysis.

At that time data processing was a new technique in constant change. Few academic lesson were devoted to this topic. Many knowledge concerning the implementation of the data processing were acquired by the practice. The design of software was the occasion to release generic concepts .

Mentioned experiences shown us the questionnaire should present a set of formal constraints to guarantee a quick and reliable treatment of the data. We gradually drew a series of general principles from the constraints we met. PRAGMA allowed us to model, to instrument and to evaluate their relevance.

The databases that were entrusted to us for statistic exploitation were generally constituted in management context and through non-directive interviews. In this case, the difference between the communication function, and the processing structuring one, is clear

because we are confronted with the organization of the treatments without controlling the conditions under which the survey proceeded.

In this case, we had to deeply restructure these bases to analyze them with quantitative and qualitative statistic techniques. We had to determine a fixed questions list from information filled about most of the individuals.

Firstly, we faced the answers absences, because it was infrequent that all the information that was eventually selected is filled-in for all the individuals.

Then, we had to study the open questions to write down a list of the answers so as to reduce this list to few modalities when the answers quality allowed it.

Then, the counts shown that some modalities of the closed questions were very marginal and had either to be eliminated or regrouped with other modalities.

Afterwards, we face questions with multiple answers that we could analyze in several ways: by considering the first answer, by breaking up the answers, by scoring them, etc.

We also had to prepare the data analyses by selecting characters among the questionnaire modalities set and by affecting a mnemonic code to each character.

### **1.3.2. PRAGMA formal principles**

Thus, we quickly decided that it was preferable that the questionnaire, as an instrument to structure data in the prospect of their treatment, be drafted and used in such way that each individual answers all the questions, all the questions are closed and each of them only accepts an answer.

It is most certainly a theoretical model that is rarely executed. That is why we speak about principles rather than about standards. This model is useful to design questionnaires, to restructure databases that are usually constituted on the base of forms that are very different from this reference, to organize treatments and to integrate them.

We should notice that the mentioned principles: compulsory answers, closed questions, sole answer, are merely formal. They concern the questions form independently from their formulation and meaning. The difficulties we face when they are not respected, as the answer absence, can reveal that a question is not formulated properly, is not understood properly...nevertheless these principles remain formal.

They became more important when we passed from diagnosis to observatories. A diagnosis is similar to a survey for which study we have enough time, even if it is reduced by

the concern to publish the results soon. In the observatories, the results should be regularly and more quickly published. Ideally, an observatory aims the real time. With Internet, this ideal is becoming reality. To reach this objective, we should model the treatments according to automated protocols. The treatments repetitive character also implies such protocols.

However, the questionnaires design shows these principles remain ideal because, on the one hand, reality cannot be confined in these standards, and on the other hand, because of communication function of the questionnaire.

In reality, the respondents can have several children, several income sources, etc. Consequently, they should be able to give several answers to a question. From the communicational point of view, there is no obligation the respondent knows every thing and has an opinion on every issue. We already explained that the communication and the structuring processing functions can lead to contradictory choices at the time to choose the question form, open or closed.

That is the reason why PRAGMA is not normative. It is not constrained at the beginning with the strict respect of these principles. PRAGMA guarantees a quick and reliable treatment if the user can respect them. Otherwise, if the user wants to treat a database that was constituted whilst ignoring these principles, PRAGMA offers all the coding functions that will allow isolating the quality-data and treating them in the best conditions.

PRAGMA should never block the user because it did not respect a principle or a standard. In this case, it must simply indicate to the user the consequences of his choice and help it to implement the process of data structuring appropriate to the desired treatment.

## ***2. QUESTIONNAIRE SPECIFICATIONS***

We have seen PRAGMA considers the questionnaire as an interactive document of information gathering that structures the latter so as to simplify its treatment. We are under the best conditions if each surveyed people fill in all the questions, in the same order and with the same formulation. We note that when the statistical individuals are not people but items, the most important constraints are the questions number and meaning. The meaning of each question is established before the data gathering. It should not be modified during the survey. It is indispensable to define a common language to all the people who will be associated to the data gathering, analysis, interpretation and publication. PRAGMA considers that the questionnaire design constitute a preliminary task. It allows digitalizing the questionnaire when it is drafted. It controls then the form of the questions but not their formulation.

PRAGMA first function is the questionnaire digitalization, according to specifications that aim at automating and integrating treatments and the data coding. The questions number, order and formulations are established during the questionnaire digitalization. It is impossible to automatically control the meaning variations. PRAGMA does not do it, but it coaches the human control of the data meaning permanence so far as it is possible. It notably helps the user identifying the meaning modifications that result from the data coding.

Then, PRAGMA will be used for the direct or deferred data key-in, from a “paper” questionnaire. Several persons generally make the key-in. When the key-in is finished up, the data are regrouped within a sole PRAGMA.

Then, one or several other persons make the treatments until the publishing of the results.

According to their functions, the various users have more or less restrictive access rights. These rights also protect the data access. They allow some people to make operations that require deep knowledge and skills and other ones to make their work neither having all the knowledge that are necessary to implement PRAGMA, nor controlling all the skills; indeed, they only have the skills that are useful for their work.

We will examine successively the rights of access, then the specifications that PRAGMA must respect at the time of the questions and modalities recording, in order to control the key-in, then to guide the implementation of the treatments.

## ***2.1. Access rights***

In its original version, PRAGMA distinguishes three types of access rights:

- Consultation.
- Key-in.
- Definition.

Beyond protecting the data access, these rights define users' profiles from the point of view of evolution. The accesses become more complex when PRAGMA is shared or when it is integrated in an information system. A shared use requires a strict management of the rights of access to the data concerning the individuals. Integration in an information system implies a management of the access to the documents. These profiles are then used initially to affect bundle of access rights that can be arranged then by a system of keys according to the users, the individuals and the documents. We will outline these evolutions, which will be detailed in the report "Specifications for the PRAGMA integration with the software of qualitative data analysis ANACONDA and NUAGE" (deliverable n° 55).

### **2.1.1. Consultation profile**

It is the minimum profile of the users who can only consult the data and the results without possibility of modifying information. In the original version there is no restriction for the consultation. In a shared system, one can limit the consultation to a restricted group. The access right can also be declined according to the information: all the users can consult public information whereas other information is accessible only to some users.

The access to a result can be limited to some users during its drafting, and then it can become public when it is validated. Here, we can see that the rights are linked on the one hand to the users, according to their relation with the document, and on the other hand to the documents or at some moments of its life.

This implies a very flexible system of management of the rights making it possible to assign in a first time to each user a general profile, then to adjust with keys the access rights of each user according to the documents.

### **2.1.2. Key-in profile**

It is the general profile of the users who can make key-in operations: adding individuals, registering answers, modifying them and possibly suppressing them. PRAGMA

was designed for the shared key-in. Consequently in principle several users benefit from this right for a same questionnaire.

This point is important in the PRAGMA evolutions, especially for an online development.

On the one hand, these users can constitute a team within a same organization. In a partnership in which several organizations cooperate, there are several teams. A same team can also participate to several partnerships.

On the other hand, we can organize the individuals that are the key-in subject into several groups within an organization or a partnership. In the simplest case, each user manages a group. The case in which a team manages a group is also simple. But one can easily imagine that a user manages individuals belonging to several groups. An individual can also have an itinerary during which several users will manage it, successively or simultaneously.

On this point as well, it is compulsory to imagine supple and evolutionary rights systems that link on the one hand the PRAGMA users who key-in information about individuals, and on the other hand these individuals that are organized into perennial or temporary groups.

### **2.1.3. Master profile**

In PRAGMA the right of definition corresponds to the management of the questionnaire. This “master” can create a questionnaire, add, modify or suppress questions, define and modify the questions properties, add, modify or suppress modalities, define or modify their properties. He can also define the groups and the access rights of the other users of this questionnaire.

The administrator can also make treatments. Skilled users can also be in charge of the treatments, according to varied rights that link these users to some treatments or to some treatment packages.

Conversely, we can easily make accessible the treatments that do not present any data corruption risk and of which results quality is controlled. In a participative logic, the users that have key-in rights can also participate to some treatment steps.

Rights, especially validation ones, can also govern the results edition.

## ***2.2. Questions specifications***

The master carries out the recording of the questions in the order of the questionnaire and according to the established formulation, and he defines their properties, according to precise formal specifications, of which condition controls, coding and the processing.

### **2.2.1. Question row**

The question row is automatically indicated. It does not necessarily correspond to the row that is indicated on the « paper » questionnaire when the questionnaire design did not foresee the constraints that are linked to the treatment. If we want to make a correspondence with the paper questionnaire, it is necessary to integrate the « paper » row at the beginning of the question formulation. Nevertheless, the row is not a perennial identifier of the information that is why a mnemonic code is preferred. In the observatories, which periodically repeat the diagnosis, questions rows change with the suppression or the insertion of questions.

### **2.2.2. Question formulation**

The question formulation is not standardized. Now, it is necessary to forecast two formulations, one under the form of questions that use a direct formulation directed to the respondent, without length limitation, for instance “What is your gender?” and the other one under an indirect form, that is limited to sixty characters for the results publication, for example “Gender”.

### **2.2.3. Question Code**

A mnemonic code identifies each question. It is composed of two letters that are not accented, a block letter and a little one, for example “Ge”.

### **2.2.4. Question form**

The question open or closed form is not defined *a priori*. It is possible to key-in the answer at the same time under the open and the closed form. PRAGMA offers at the same time a frame in which the respondent can freely express his answer with a usual word processor and a modalities list the respondent will be able to tick by a click. In principle, and if the question does not include modalities, it is open. We will be able to add the modalities as long as we are coding the answers. It remains possible to permanently use the two answers modes, especially to specify modality content. In this case, PRAGMA does not establish any link between the concerned modality and the answer that specifies its content, except if the user specifies it in the answer content.

Then, the answers coding allows making evolve the open questions into closed ones. In this way, one can prefer an open question for the survey, then move to a closed form adapted to the treatment. Of course, it is preferable to envisage this work during the questionnaire design, rather than to discover it during its processing.

### **2.2.5. Modality without answer**

Each question automatically include a modality “without answer” (of row 0, coded “wa”) for the case in which the respondent did not indicate an answer. In this case, the user must click this modality. If no answer is indicated during the key-in, PRAGMA automatically records the modality “without answer”. At the moment of recording one individual, PRAGMA controls that all the questions have an answer and it automatically records a modality “without answer” each time it is not the case. The “without answer” are automatically entered.

### **2.2.6. Modality without object**

Each question automatically includes a modality “without object” (of row “X”, coded “wo”).

When the question is possibly without object the master keys in a “without object list” the modalities mnemonic codes that provoke the question skipping.

For example the respondent cannot indicate the number of children he has if he does not have children. In this case, the master will key-in in the “without object list” for the question “How many children do you have?” the mnemonic code of the modality “No” to the question “Do you have children?”. So PRAGMA will skip over the question “How many children do you have?” each time the respondent will answer “No” to the question “Do you have children?”

As conditions can be more complex than in the previous example, this list has a table form. The modalities of which codes are written on a same line should all be filled-in to provoke the question skipping (relation “and”), whereas it is sufficient that one of the modalities, or modalities association, of which code or the association of codes is on one of the different lines, be filled-in to provoke the skipping (relation “or”).

In this way, the question skipping is defined as a property of the question (B) that can become without object, and not of the (A) question that contains the modality of answer that makes this question (B) without object. It is possible to cause the jump starting from the question (A) of which answer can cause the jump, but in this case, the question(s) (B) without

object must follow the latter directly. It is not always the case, this is why it is preferable that the jump be defined as a property of the questions (B) being able to become without object.

Of course, that implies that the question (A), which can cause the jump, appears in the questionnaire before the question (B), which becomes without object according to the answer to question (A). It is advisable to check it during the questionnaire design.

### **2.2.7. Question with unique or multiple answer**

A property of the question describes if the latter admits one or several answers. It allows controlling the key-in afterwards.

## **2.3. Specifications of the modalities**

The master keys in the modalities in a specific list, after the modalities “without object” and “without answer”.

### **2.3.1. Modality row**

A digital code corresponding to their row in the list is automatically attributed in the order of their key-in. In the old versions of PRAGMA, it is this digit that is registered as answer when the modality is clicked. Now, it is necessary to prefer the mnemonic code, which does not change when the row of the method changes.

### **2.3.2. Modality mnemonic code**

The modality is also identified by a mnemonic code, which includes four non-accented letters. The first two letters correspond to the question code, the following two ones, a block letter and a little one, more precisely identify the modality – for instance “GeFe” for the modality “Female” of the question “What is your gender”.

PRAGMA allows adding, suppressing, inserting or modifying modalities at any moment, and even during the key-in. In this case, a modality row, and consequently its encrypted code, can change, what is not the case of the mnemonic code that is linked to it. That is why the PRAGMA new versions do not register the modality encrypted code (the row) but its mnemonic code that does not change, even if the modality row changes.

A modality code is unique. The PRAGMA new versions will have to check the unicity of a modality code when it is created, whatever is the creation mode: by direct adding to the list, by answer coding, by recoding or by formulation modification.

The data export and import operations also use the mnemonic codes. Thus, for example, the “GeFe” code, that comes from digitalized questionnaire that was digitalized in

PRAGMA will be always re-imported properly in another questionnaire, whatever is the row of the question “What is your gender?” in the questionnaire and the row of the modality “Female” in this question”, provided that the code “GeFe” remains reserved to the identification of the modality “Female gender”.

When a modality is selected to constitute a character for a data analysis, the character keeps the modality code to which it corresponds and the presence of the question code allows quickly identifying the question from which the character comes.

### **2.3.3. Modality Type**

PRAGMA dissociates three modalities types according to the question type.

The modality is “exclusive” (type “E”) if it corresponds to an answer that excludes the other modalities. For instance, the modality “I do not know” excludes any other answer. All the modalities of a question that only admits an answer are exclusive. For example, the modalities “Male” and “Female” are exclusive one from the other. Only one answer can be key-in because the choice “Male” erases and replaces “Female” and *vice versa*.

The modality is “unique” (type “U”) if the question admits several answers but this modality can only be answered once. If we try to click several times the modality, it will only be registered once.

The modality is “multiple” (type “M”) if the question admits several answers and this modality can be answered several times. For instance, if I have two children, they can both belong to the same age class. Or several members of a same family can belong to the same socio-economic group.

Whereas a closed question with unique answer only includes exclusive modalities, a multiple-choice question can include the three modalities kinds.

### **2.3.4. Modality frequency**

PRAGMA calculates in real time and it also displays a column “Frequency” that totals up the numbers of individuals who chose each of the modalities. Thus, any person who keys-in a questionnaires file has modalities frequencies in real time.

### **3. SPECIFICATIONS CONCERNING DATA**

The answers form is more important than the questions form as regards processing. We can distinguish three, and only three, answers types:

- The code, including Boolean code;
- The measure characterized by an order property;
- The text.

The answer code corresponds to the closed question. The code is the unique element that allows directly and automatically making simple or cross sorting, and consequently count tables, and then histograms as well as maps. It can be then automatically break up in characters for qualitative data analysis. In this case, the answer is a Boolean code « 1 » or « 0 », depending on whether the individual owns or not a reference character.

The measure and the text are answers to an open question. Answers to the open questions are called “values” in PRAGMA. Values should be coded before performing the elementary statistical processing that is the quantitative sorting. It is a relatively simple operation for measures that have a property of order. Coding is longer for texts and it depends on the data quality.

Before observing the specifications concerning codes, then values coding, we will specify those that refer to the individual management by PRAGMA.

#### **3.1. Individual**

The individual, more exactly the statistical individual, is the basic element of a statistical processing or a data analysis. Within a survey, the individual is a human person, who answers in writing or in oral to the questionnaire. But the statistical individual can also refer to objects or other living beings than a human person, that are observed with the help of a questionnaire, that is to say by collecting the same information – the same number of information – for all these objects or beings, in the same order and with the same meaning – which depends on the observer.

Whoever or whatever is the statistical individual, PRAGMA allows gathering the same number of information for all the individuals, in the same order, with the same meaning. Each individual is then described with the same information, the questions. Only values of these information change from an individual to another, the answers.

A survey is defined as an individuals set, who distinguish themselves separately with their answers profile. Formally, each individual is identified by a code and it consists in an answers set, which can be identified to an information vector, or a form. It is the unitary element of the survey. We consider that these individuals are statistically separate, that means that we cannot deduce the answers of an individual from those of another or many other individuals.

Things are not always that simple. A questionnaire can refer many individuals. For instance, if we question persons about their children with specific questions about every child, we consider in reality two types of individuals: persons in one hand and children on the other hand. The diagnosis and assessment guide then refer persons and households. PRAGMA does not enter into this type of consideration; it considers that the individual conventionally is either the person or the household, according to the statistical individual to whom the questionnaire is addressed. If the questionnaire addresses itself to household, as assessment guides about the Minimum Insertion Income in France, the individual is the household, even if it comprises specific questions for persons forming the household. At present, the diagnosis and assessment guide, which addresses itself to the person, comprises questions about the household. We note that we should pay attention to specify to which entity – person or household– addresses itself a question – such as the income for example, because most of questions address themselves to the individual, and some others to the household, even to other persons as spouse or children.

The time intervention also complicates the precise resoluteness of the individual. Within the strategies of follow-up or observation, a same person could be questioned many times, within the framework of a path or a sample group. When we question a person for the second time, do her/his answers substitute or complete the answers that he/she gave the first time, or do I store them as if it was a new person? In this last case, I will have several observations for each person, and the statistical individual, the processing unity, is the observation.

At the survey level, the individual can be formalized, as a vector constituted of an identification code and an ordered list of answers. Then, it corresponds to a line of a data table, or a form in a file. We refer to the notion of data table but we know that the technical solutions concerning databases can be very different. The data table is a simple notion. We often transfer the results of a small survey on a big paper sheet by entering the individual answers on a line, within columns corresponding to different questions. This practice gets

generalized with spreadsheets like Excel. The data table shows within each line the individual answers, an answer per column in the questionnaire order.

### **3.1.1. Individual identification**

PRAGMA identifies individuals with the help of a group code and a distinctive number.

#### **3.1.1.1. Group**

The group notion is acquired from the shared collecting. Once the questionnaire established, the PRAGMA software was spread between many data entry points. The group was a simple manner to identify each of these points. Within each point, each individual was identified with a number. The group referred to an organism or a territory, it makes no difference. Once individuals gathered, it allowed identifying the data entry source.

A code with three letters identifies each group. Groups are established according to an organization principle (organisms, actions, users, territories, etc.) at the discretion of the questionnaire supervisor, who assures the organization and management of the shared data entry. The letters allow establishing mnemonic codes.

#### **3.1.1.2. Distinctive number**

It is a number with eight figures, which is automatically assigned, in the data entry order, to every new individual. It is unique within the group. If an individual is deleted, the individual number cannot be assigned to another individual.

In terms of a shared data entry, it does not always exist an interconnexion, which would allow controlling that the number assigned to each individual is unique. However, if each data entry point is well identified with a unique group code, the number will be unique within this group. Individuals appertaining to different groups can have the same number, but they have a unique identifier with the association of group code and distinctive number, for instance PRA00000001.

#### **3.1.1.3. Observation number**

When an individual is subject to several observations at different time, it is the observation that becomes the real statistical individual. Therefore, we can keep the reference to the person or the object that the individual would constitute in case that only one observation would be made. In this case the unique figured code is assigned to the observed

person and it is completed with a two-figures number corresponding to the observation row and which one have to be unique, even if the observation is after-deleted.

#### **3.1.1.4. Anonymity**

When the survey is about persons, what is the case for CATALYSE diagnosis, the identification code of the person is perfectly anonymous, even at the level of the questionnaire supervisor.

Within physics, even within life sciences, statistical analysis do not usually interest in statistical individuals. Their main quality is the neutrality. If we should assign a unique identifier to each of them in order to distinguish them, there is no interest in identifying them individually.

The situation is different within human and social sciences, in which researchers and actors are often more interested by individuals than variables. Archeologist wants to identify excavated buildings, geographer wants to identify his/her landscapes and most of actors want to identify persons of whom they are in charge.

The principle of PRAGMA is that only the person who stores data can establish a link between an identification code and an individual. He/she is the only one able to correct data in case of mistake. He/she is also the only one able to establish a link between codes and individuals of which he/she is in charge during the results interpretation.

#### **3.1.2. Individuals and data tables types**

We can dissociate three types of individuals and data tables contingent on data types, which constitute the answer vector of this individual: values, codes or Boolean codes.

Data table is a table composed of lines and columns. It can be easily edited in the form of a spreadsheet. Whatever is the type of table, lines correspond to individuals. The first column is reserved to the individual identification code (it will be excluded later when we will talk about columns). The following columns contain answers, in the same order for every individual. Columns correspond to questions according to a logic of questionnaires processing, then to characters according to a logic of data analysis. Table columns are separate: a column cannot be inferred from another one or from the combining of many others.

We assign as « unrefined » individual, an individual of whom some answers are values, measures or texts. It can combine all types of answers.

Columns of an « unrefined data table » are questions whose some at least are open. In other words, some columns contain values. Boxes concerning a column will incorporate a value if the question allows only one answer, or many – isolated by a separator sign – when the question allow many answers. It is the table that we obtain from a survey including open questions or more generally from a file or a database, which has been constituted with a management finality and not with a statistics analysis finality. This table should be coded in order to be subject of statistical and spatial processing.

An individual is « coded » when all his/her answers are codes.

« Codes table » columns are all closed questions. The codes table lend themselves to statistical sorting insofar as questions are with unique answer. It can be used to map if one of the questions allows a geographical reference.

An individual will be called « Boolean » if his/her answers are Boolean codes, “0” and “1”. In this case, codes do not refer to questions but to characters, which correspond to modalities of closed questions.

A « Boolean table » exclusively contains Boolean codes. Its columns are now characters. It is the table type used by the data analysis. We can automatically establish a Boolean table by splitting a codes table. A columns number equal to the modalities number that this question contains substitutes each codes table column, which corresponds to a question. Modalities are then called characters and the answer is indicated with an “1” within the column corresponding to the possessed character, and “0” within the other columns. In case of multiple answers, we will indicate much as “1” as possessed characters. There cannot be many answers in a box of the table.

The progression of the unrefined table to the Boolean table is the reference of processing advised by PRAGMA. In particular, whatever is the chosen technical solution for the databases management, the data table in the form of text format file editable with a spreadsheet, constitutes the exportation and importation standard of data. The typographical character “+” constitutes the separator of multiple answers.

### **3.2. Code**

The code merely allows identifying the answer in an economic way. We key-in it more quickly, it uses less space in the memory and it accelerates the sorting. Nevertheless, event if the code is a figure or a number, it corresponds neither to a filing nor to a measure.

Even if it is ambiguous, the use of coding “1” the male gender and “2” the female gender corresponds neither to a filing nor to a measure. The code only allows identifying the two codes by means of a unique typographical mark. We can also use letters, a and b, or mnemonic marks, as m and f.

The rows of questions and modalities are traditionally used to encrypt them in the questionnaires, because digital codes are very practical to programme the quantitative sorting.

PRAGMA prefers the mnemonic codes used in data analysis because they identify in a surer and more permanent way the questions and the modalities.

### **3.2.1. Digital code**

Historically, each question code is a number that corresponds to its row in the questionnaire and each modality code is a number that corresponds to this modality row in the list of the suggested modalities for the question.

Nevertheless, the questions modalities are not clearly identified by their row within the question. On the contrary, all the row 1 modalities are coded “1”. To clearly identify a modality, it is necessary to refer to it through a numbers couple of which the first one is the question row and the second one is the modality in this question. Thus “2,2” refers to the female gender if the question “What is you gender?” is the second in the questionnaire and the modality “Female” the second in the modalities list. Moreover, if the questionnaire evolves during this time, the codes of some questions or modalities change when we insert or suppress a question or a modality.

Or, in an observatory, it is logical that the questionnaire evolves during this time, we suppress questions, we add others as well as modalities.

### **3.2.2. Mnemonic codes**

That is why PRAGMA presently prefers mnemonic codes.

The mnemonic codes are codes that are composed of letters and numbers. The possibility to use several letters or numbers allows making a code that evokes the modality meaning, for example “Fema” for the female gender. The mnemonic codes use is common in data qualitative analysis. In this case, we do not speak about modality anymore but about characters. Nevertheless, it is the same thing for each modality, which can become a character.

PRAGMA uses mnemonic codes that are composed of four non-accented letters and possibly of numbers, except other typographical marks, without distinction of block letters and little ones. The four-letter limit guarantees all the data analysis software can read them.

Because of a convention that is proper to PRAGMA, the questions are coded with two letters, except numbers, mentioning the question then the modality. Modalities and characters are coded when the question code is completed by a specific code that is composed of two letters; for instance “GeFe” for the modality “Female” of the question “What is your gender?” and for the character “Female gender”. We make easier the code reading with the use of a block letter and a little one. It is a convention because PRAGMA does not distinguish block letters and little ones, as many operating systems or statistics analysis software. Many software of data analysis use mnemonic codes made of four characters. We advise this standard even if PRAGMA can manage longer codes. It can use digit when the modalities are ordinates. However, it is not recommended if new modalities would be introduced. It is thus advisable to be very careful in the use of digit.

We remind that the code of each question, of each modality or of each character is unique. It allows identifying it without any ambiguity. PRAGMA checks the code uniqueness when it is created. In case of recoding, which can be made with the gathering of several modalities (or characters) into only one. This new modality does not have the same meaning as the gathered modalities. It must allow it to have a new code, different from those of the gathered modalities.

### **3.2.3. Boolean codes**

For your information, the Boolean code is either the figure « 0 », or the figure “1”. It shows the answer in terms of possession of a reference character. This character is identified by a mnemonic code. The Boolean code is used for data analysis.

### ***3.3. Studying and coding values***

In order to code an open question, it is advisable to study the question first in order to provide the list of values.

We calculate the frequencies of each value during the same operation. These frequencies are usually disparate; some values are well represented, whereas the others are very unconventional and thus non-representative within a statistical analysis.

The coding differently occurs depending on whether the values are measures or texts. The measures are divided into classes independent of their meaning, whereas texts are gathered into categories depending on their meaning. Classes and categories form modalities at which codes are assigned.

However, PRAGMA proposes an only one coding function whatever is the type of value, measure or text. The measures are simply arranged during the count. The difference is thus only showed here in a theoretical way, for future evolutions.

### **3.3.1. Measures division into classes**

The measure owns a property order that allows filing the answers without knowing their meaning. We all know that 15 are inferior to 20. It is not important at all to know if the topic is the age or the weight. Open answers to questions « How old are you? » or « What is your birth date? » are examples of measure.

Thus, it is possible to arrange the answers list from the minimal value, or minimum one, to the maximal value, or maximum one.

The dividing changes measures into classes, which ones can be coded like modalities. We just need to specify the lower thresholds, or the upper ones, of classes. A general function automatically provides the dividing into classes and assigns to each measure the code of the corresponding class. The minimum constitutes the lower threshold of the lower class. The maximum constitutes the higher threshold of the upper class.

The dividing can have three objectives. The first two ones are separate from the question meaning and automatically performed. In the first case, we want to constitute classes of equal interval, and in the second case, classes of equal size. We just need to indicate the wished number of classes in order that the thresholds are automatically determined. The third case consists in keeping significant thresholds, for instance 18 years old for the major age; it involves a human decision and possibly the intervention of an expert.

In this last case, it is possible to store the thresholds as parameters for a future use, especially if their purpose need an expert intervention, or more simply if we wish repeating the dividing. An untrained user will use these parameters so that the dividing of the question will seem to him / her to be an automatic operation, we will say transparent in this case. The setting is also used to repeat the dividing through time and distance, in order to compare diagnostics realized on different territories or establish temporal streaks in the observatories.

This type of setting constitutes an issue within the definition of the CATALYSE toolkit to propose, thanks to settings, protocols of transparent processing for a neophyte user. A more experienced user could modify the parameters as he/ she pleases.

### **3.3.2. Gathering texts into categories**

Every answer to an open question, which does not have order structure, is a “text” answer. It can consist of one or many words. The question “What is your job?” is an example.

As the answers do not a priori have a formal structure, values are alphabetically categorized and the definition of categories is executed according to the answers meaning. The alphabetical order has only here a practical function. A “Secretary”, a “Scientist” and a “Seaman” do not belong to the same category because they are close within an alphabetical list or they start with the letter “S”.

PRAGMA studies answers with the help of a disassociation or an association of multiple answers. It does not operate a lexical dividing or a content analysis, but it is possible to export the values in order to make it with appropriate software.

The coding through gathering is based on the values meaning, it consequently involves a high quality level of data. It associates values of which the meaning owns a certain precision level, within a category, of which the meaning is more general, less precise. For example, “Metal worker” and “Milling machine operator” will be gathered within the category “Worker”. If answers include at the same time values with different precision levels of meaning, such as “Metal worker”, “Milling machine operator” and “Worker”, the relevance of the open question is arguable because the gathering must anyway happen at the least precise level, here “Worker”. We also have to pay attention to homonyms, like “Watchmaker”, which can assign a “Worker”, but also a “Shopkeeper”. The choice of an open form must consequently be made with the consciousness of this demand in the instance of texts. Otherwise, the cost of a bigger gathering as well as the coding will not be justified, especially if the question is eventually not exploitable.

The partial automation of gathering can be made with the help of a dictionary, which will punctually need a human intervention to each new value or ... misspelled value.

#### ***4. SPECIFICATIONS OF CONTROLS FOR ROBUSTNESS AND DATA QUALITY.***

PRAGMA integrates controls procedures that prevent errors of the less experimented users and that more globally aim at guaranteeing the data quality. These procedures objectives are:

- Giving security to the information gathering in relation with the handling errors.
- Improving the shared gathering quality.
- Allowing the restructuring of the data bases so as they lend themselves to a quantitative and qualitative statistical exploitation.
- Guaranteeing quick and reliable treatments.
- Allowing a cooperative exploitation.

These procedures concern:

- The automatic controls that particularly prevent the handling errors.
- Coding functions that allow improving the data quality or that allow giving the appropriate form to the contemplated treatments.

##### ***4.1. Automatic controls***

The automatic controls are about:

- Answers absence.
- If the question allows only one answer or many answers.

##### **4.1.1. Answers absence and question without object**

PRAGMA locates and systematically accounts the answer absence by automatically dissociating situations in which the question becomes without object.

The main constraint of questionnaire survey that is for each question to be filled-in by all the individuals is rarely respected. We frequently face the problem of answer absence. In this case, it has no importance that the surveyed person or the observer could not answer, did not want to answer or forget doing it; the fact is the answer is missing, imprecise or unreadable. It cannot be clearly determined. It is absolutely forbidden to interpret an absent answer or an answer impossible to determine. PRAGMA allows the user, who makes the data entry, to indicate the modality “without answer”. We saw that when the user misses informing

an answer during data entry or wants to store an incomplete questionnaire, PRAGMA automatically completes each non-informed question with the code of the modality “without answer”.

This case should not be confounded with the situation in which the question becomes irrelevant. We saw that in this situation, there is a logical reason for the answer absence established by an answer to an anterior question. We also saw that PRAGMA allows defining for each question the cases in which it becomes without object. During key-in, it automatically skipped over the questions without object and filled them with the code of the modality “without object”.

The situations in which the question is without answer and those in which it is without object are separately counted, what allows a precise calculation of the missing answers frequency.

The missing answers complicate the sorting and results analysis because the total number of answers is under the number of surveyed individuals. It is possible to calculate two percentages: in relation with the individuals and according to the number of expressed answers. These percentages are quantitatively different and they do not have the same meaning, what generates confusion risks for a user who is not very well skilled. It is a reason for avoiding the answers absences.

In the instance of questions without object, it is yet possible to restore the total number of expressed answers at the level of the surveyed individuals number, by completing the frequencies of modalities with the one of the modality, which causes the question skipping. This operation leads to synthesize two questions (see below section 5.5).

#### **4.1.2. Unique answer and multiple answers**

PRAGMA also precisely checks the keying in of the questions with “only one answer possible” or with “multiple answers authorized”, and especially if the modalities are exclusive, unique or multiple.

When the question includes and allows several answers, they are registered in the order in which they are keyed-in.

The questions with multiple answers are also more complex to sort and to analyze because the total number of answers is more important than the number of surveyed individuals. It is also possible to calculate two percentages, what also generates a confusion

risks. That is why the questions with unique answer are preferable. Nevertheless, there are many concrete situations in which we should use questions with multiple choices, for instance because the households can have several dependent children. In this case, it is possible to replace a question with multiple answers by as many questions with unique answer “yes” or “no” as modalities. If it is too difficult to manage as regards communication it will be more appropriate to automatically change the question before the processing.

The systematic controls of answers absence, of the situations in which the question becomes irrelevant, of the status of the question and of its modalities according to the fact they admit a unique or several answers, as well as handling errors at the data-processing level, make PRAGMA a robust tool. It could be used in extremely diversified conditions, very close to the observation situations, by people who do not have a data-processing or statistic qualification –what does not mean without formation and follow-up-. In exchange, these experiments allowed improving PRAGMA robustness, which is an important factor of conviviality.

#### ***4.2. Coding and recoding functions***

It is not possible to automatically control every thing. As we explained before, PRAGMA is not a normative tool that excludes the individuals who did not answer all the questions and what obliges to only use closed questions with a unique answer. It allows on the contrary importing information from databases, which have not been developed according to this formal model that incorporates constraints of quantitative and qualitative statistical processing.

PRAGMA completes the automatic controls with coding and recoding functions, which structure data contingent on the wished processing, following a protocol that aims to automate the tasks of sorting, count, calculation and results representation in order to restrict human interventions to the decision-making phases.

The automatic controls then provide a time saving about sorting and calculations, which are tedious operations, not very attractive and sources of mistakes. They prepare the effective implementation of this protocol, which aims to improve the data quality by adapting their form to the processing to perform. Then, they certify the respect of the formal specifications about questions and answers during the protocol, and they prepare the decisions while preempting possible mistakes.

For instance, the control of missing answers identifies these last ones, separates them from situations without object and calculates their frequency. It prepares the decisions concerning data elimination, of which relevance or meaning does not seem to be understood by everyone, or of which representativeness is put in doubt by an important frequency of answers absences.

In this way, the values coding and the recoding aim to improve the representativeness of modalities for a quantitative processing, then characters for a data qualitative analysis. The decisions about the coding and recoding involve the elimination of answers or aim to gather non-representative answers. They should respect the meaning of answers. The gathering also modifies the meaning of answers.

Thus, the decisions are based on two criteria of data quality: the representativeness and the meaning, one quantitative and the other qualitative. They often use these criteria together.

The representativeness of answers is a fundamental criterion of data quality for a statistical analysis. The notion of representativeness is linked to the one of statistical cross-section. The statistics showed that it is not indispensable to question all the individuals, who constitute a population, to determine the general attributes of this population. Beyond a certain cross-section size, the information gain obtained thanks to the questioning of a bigger number of individuals, is reduced compared with the cost of the data gathering and data processing. Conversely, we must question enough individuals in order that the results observed or inferred with the cross-section analysis are transposable to the whole population with an insignificant risk of mistake. This constraint also establishes itself to each modality, that the set of surveyed individuals, who own it, constitute a subgroup of individuals who own it within the observed population. Even when the set of surveyed individuals is not composed as a cross-section, each modality has to own a frequency enough important to be inferred or planned on the basis of a quantitative or qualitative statistical analysis. PRAGMA does not delimit the representativeness threshold of modalities, which must be independently calculated. When this threshold is delimited, for instance a frequency equal or higher to 5 % of the surveyed individuals with an absolute minimum of 15 individuals, PRAGMA could indicate the answers that are not representative and that should be a priori deleted. This suppression, that manifests itself with a loss of information and missing answers, can be avoided thanks to the gathering of answers into only one whose frequency will be upper to the representativeness threshold, on condition that the meaning of gathered answers allow it.

The meaning – a consistent understanding of the data meaning for all those who are involved within the analysis – is also a fundamental criterion of data quality. A high frequency of missing answers frequently reveals an understanding problem about the question. The gathering of values or modalities during the coding or recoding of data aims to constitute representative modalities but it should mix values or modalities together, which share elements of meaning. The coding and the recoding must also be in charge of the good understanding of data and results for all those who will have to analyze them, explain them or simply read them, within workshops more or less limited to collaborative work space or Internet.

The protocol of data qualification then replies to several finalities. Except for the selection of representative data whose meaning is clear for everybody, it allows selecting questions or modalities according to the wished analysis stage, which can be exploratory or advanced. They are also useful to prepare the results communication. The report “Specifications for the PRAGMA integration with the software of qualitative data analysis ANACONDA and NUAGE” (deliverable n° 55) and the paper “The editorial function of the territorial intelligence systems” [GIRARDOT, 2006] began the inventory of functions concerning the information processing and the description of analytical and editorial protocols within the territorial intelligence systems. We will get here onto the functions concerning PRAGMA.

The PRAGMA qualification and coding protocol organize four steps:

1. Control of the missing answers;
2. Coding of the open questions;
3. Statistical recoding of the questions;
5. Selection of the characters in the prospect of the qualitative analysis.

The functions, which allow the accomplishment of these stages, are localized in the screen-window of each section, because every question does not need the same processing. It is also possible to activate them from the questionnaire stage by planning to determine and/or select automatically the questions concerned by the function.

#### **4.2.1. Control of the missing answers**

The control of the missing answers concerns at the same time the individuals and the questions. In principle, we should eliminate all the individuals for which an answer or more is

missing. Nevertheless, the data acquisition cost or the concern in not introducing bias in a sample can oppose to a so drastic information loss. In this case, it is necessary to eliminate the individuals and the questions for which the rate of missing answers is too important, for example higher to 5%.

We record that we should beforehand deduct from the missing answers the situations in which the answer is irrelevant to count the missing answers. That is the reason why it is recommended to identify the questions for which these situations can take place during the keying-in of the questionnaire, especially the questions that could be without object.

Thus PRAGMA can automatically determine these situations and directly count them on the screen-window of each question and in the assessment report, which can be edited and printed with the help of a spreadsheet.

Insofar as the control of the missing answers leads to eliminate individuals and questions, it is the first step of the protocol of data qualification. At the end of this step, as well as at each step, the balance could be easily updated to control the results of the processing and to prepare next steps.

The operations that the user could work on are:

- The suppression of questions for which the number and the rate of individuals who did not answer, is too high;
- The suppression of individuals who did not answer to a certain number or rate of questions.

PRAGMA do not delete questions or individuals. In a general manner, the suppression is an irreversible operation, which should at least be confirmed by the user when it is inevitable, in order to avoid the consequences of a takedown error. It is preferable that an external administrator confirms it before it becomes definitive.

PRAGMA deselects the elements that the user wishes deleting, because what have been deselected can be selected again. The operation is transparent for the user because PRAGMA acts as if the deselected questions were deleted. The selected questions begin with a uppercase, for instance « Li » for the question « What do you look like? », whereas the deselected questions begin with a lower case, for instance « li ». It is the same for individuals (and modalities).

Therefore, PRAGMA does not distinguish at the moment the deselected elements according to the stage of which they have been moved away following processing. At the end of the protocol, it is not possible to determine with security at which stage, thus for which reason, an element has been deselected. The definition of processing protocols that we begin here should allow a best memory and a best traceability about resolutions.

The coding, the recoding and the selection of characters can involve the suppression of answers. Then they make missing answers. The control of missing answers has consequently to be repeated after each of these operations.

#### **4.2.2. Coding values**

The open questions coding precedes the steps that are made on closed questions. The coding aims at changing the open questions into closed questions, that is to say the values, measures or texts into codes. It is an indispensable operation that will allow making statistic sorting on the data, and from these quantitative counts making histograms or possibly geographical maps. It prepares the selection of characters for a qualitative analysis.

As we previously explained, we can distinguish two coding forms. The cutting changes the measures into classes from thresholds. The text gathering constitutes categories.

Insofar as a cutting is also a regrouping, PRAGMA simply offers gathering values, then changing each obtained regrouping into a new modality. It is possible to code directly as modality a value whose frequency is upper to the threshold of representativeness. Then, the gatherings are about values which are not representative and/or whose meanings are close. We have to delete non-representative values or gather them with other values in order to constitute a representative group. It is not possible to gather values between them if they are not an element of common meaning. Each gathering has a more general meaning than the one of values it amalgamates. This gathering, and the corresponding modality, will have a different wording from those with mixed values.

The function « Listing » of PRAGMA provides the list of values for one question, by separating the different values, which constitute a multiple answer. It is also possible to concatenate them if we consider the multiple answer as an association, thus as a single answer. It also counts how many individuals – different or not – replied this value. It also indexes these individuals. The index allows consulting or confronting the individuals who gave a same answer to a determinate question. PRAGMA should indicate the non-representative values.

By default, this list is implemented according to the alphabetical order. The measures are listed according to the numerical order after an automatic recognition or upon the request of the user. It is thus possible to gather them into categories.

It is possible to gather values by assigning them a common generic wording. PRAGMA deletes real-time the gathered values and calculates the gathering frequency in order to assess its representativeness. The user can transform each gathering into a modality, which has the same wording by assigning it a non-existent mnemonic code. The gathered values are deleted. It would be more interesting to keep the coding trace. The values that are not gathered are preserved. The user can delete them or keep them.

PRAGMA also allows directly coding a value which meaning already corresponds to an existing modality. When we key-in, it is possible to directly code an answer or to indicate its value if we have a doubt about its allocation to a modality. Then it can be coded during the coding step, either by allocating it to an existing modality, or by creating a new modality.

When the coding is completed, the coded question can be sorted with the function « Sorting ». It is rather a control because the sorting happens real-time during the transformation of values into modalities. It is automatically added to the assessment report that does not refer to the open questions for which it only counts the missing answers.

### **4.2.3. Recoding**

The recoding is a similar operation of gathering, which is neither about values, but about codes corresponding to modalities. Thus, it interests close questions. It concerns modalities whose frequency is too low to be representative. As previously about values, these modalities are either deleted or gathered with modalities whose meaning is close. A modality obtained by the gathering of two or several modalities, has not the same meaning as mixed modalities. Its wording, as well as its mnemonic code are then different from wordings, and codes of these modalities.

The present functioning of PRAGMA is still impressed by the original conception of PRAGMA, which is now mainly aimed toward the characters selection in order to constitute and export the Boolean standard input data table for data analysis software including ANACONDA. The recoding and the characters selection still often constitute in a practical way only one stage, during which we select some representative modalities as character, whereas we move away others modalities and we gather modalities to constitute representative characters. The preserved modalities are selected, their code, which become the

one of a character begin with an uppercase. To move a modality away, we can remove its code or deselect it, its code then begins with a lower case and it will not take part in the conception of the data table. To gather one or several modalities, we just have to replace theirs codes with the same one.

This proceeding can be preserved for the recoding of modalities:

- Deselecting of modalities that we must delete, because they are not representative and their meaning do not allow gathering them with one or many other modalities.

- Assignment of a common code for the modalities that we want to gather. The new modality resulting from the gathering has a new meaning, which corresponds to the meaning part common with the gathered modalities, and which is of a more general level. Its wording is then specific and one of the wording of the gathered modalities cannot be reused. The code, which is both the tool of gathering and the identifier of the new modality, is a new specific code.

- The preserved modalities are still selected.

PRAGMA does not represent at the moment the changing of the question following the recoding, what should be performed by means of questions known as “of synthesis” (cf. supra), by creating a new screen-window for the question, following the entered question in order to represent the question such as it results from the recoding. For that, we will presently use a spreadsheet to extract the question from the assessment report, and then we will modify the count table. Therefore, we have to pay attention because the calculation of frequencies depends on the type of the question. Especially, in the case of questions with multiple answers, when we use Boolean algebra to establish a data table, 1 plus 1 is not equal to two but one. In other words, the gathering frequency can be lower to the total of gathered modalities frequencies. It is why; it would be simpler to introduce a new question, by modifying the wordings and codes of elements – question or modalities – whose recoding made the meaning progress.

Indeed, the PRAGMA use within the framework of territorial intelligence systems separates this function of recoding from the characters selection and plans its decomposition into many stages according to different criteria and aims.

The data quantitative analysis developed within the partnerial and participative observatories in interaction with the qualitative analysis. The operational group wants a synthesized quantitative photography. The evaluators want to follow some indicators. To go

their theme into detail, workshops solicit sorting about a part of the surveyed population, a group a priori delimited by one modality or the association of several modalities corresponding to a profile, a class a posteriori coming from the typology, etc. The project leaders want some other data retrievals or crossings to argue the projects, etc.

The improvement of the modalities representativeness does not constitute the only aim of recoding anymore. This one is now often motivated by the will to effectuate comprehensible tables by various publics. The communication aims became also diversified with useful synthesis to prepare the decisions of the operational group, a more detailed vision of results for the work of workshops or cooperative spaces, accessible tables for public presentations and publication on Internet. The questions are subject to many recordings of which it would be necessary to view rapidly the results and keep a trace.

#### **4.2.4. Selecting characters for qualitative data analysis**

As we have just mentioned, the coding and recoding functions of PRAGMA questions and modalities were especially developed to select the characters that will constitute the Boolean data table in the prospect of the analysis of qualitative data that will be made by means of ANACONDA. In this logical way, one first globally select the questions to be analyzed, then code the open questions in order to select the characters from the modalities, possibly by regrouping several modalities.

The characters selection has to be performed now as a distinct function, following the coding and recoding operations that can also satisfy aims related to quantitative processing or to the results communication before or after having contributed to the making of data table for the qualitative analysis.

It should also become a simple selection of characters among the recoded modalities.

The experience also shows that we perform several data analysis, the first one having more and more a function of assessment about decisions beforehand made during the coding and recoding and an exploratory function.

While now, every new coding, recoding or characters selection delete the previous one, we should at present keep a trace of the initial coding, codings, recodings and characters selections.

## **5. OTHER DATA ANALYSYS AND DATA PROCESSING SPECIFICATIONS**

PRAGMA is a software developed to make simple or cross sorting presented in count tables and to constitute Boolean data tables necessary to data analysis.

Simple sorting is realized in real time during the data gathering and during every step of the data qualification process.

The global quantitative assessment report of PRAGMA exports every count table of the different questions, following the order of questions gathering.

It is developed according to the standard « text file » which uses only two characters of formatting: the « RETURN » (or LINE FEED, ASCII 13) at the end of each line and the « TAB » (Tabulation, ASCII 11) to separate columns. This file can be read and edited with a word processing or, better, with a spreadsheet. This last one allows the presentation and the contents of tables to evolve for needs concerning the analysis or the communication. It also allows drawing automatically graphs. We can draw up maps with a cartography software of which free versions exist.

The PRAGMA philosophy is not to manifold processing options or results presentation. On the contrary, PRAGMA exports unrefined data and results through a standard file « text file », which is universally accepted. Then, they can be used by specialized software. We privilege the tools of office automation that many users employ everyday.

Cross sorting can be obtained in two ways. They can be exported into a « text file ». We can also follow the individuals selection according to a simple or complex profile and use the real-time frequencies calculation for this profile.

The data table also exports through the form of « text file », the result of the characters selection.

The integration of PRAGMA, with data analysis software within information system, possibly online, makes upraise new needs within PRAGMA.

The necessity to introduce questions known as « of synthesis », in order to constitute automatically a question from many others, but also to keep the trace of every coding and recoding data operation, constitute an autonomous progress of PRAGMA.

We will consider in conclusion some other planned progresses, as the possibility to manage the conditional repeating of every question or to use model questions, because they

are progresses, which have to incorporate themselves into the broader prospect of PRAGMA integration with the analysis software within the framework of a territorial intelligence system.

### ***5.1 Flat sorting***

Within PRAGMA, each question is automatically subject to a simple sorting, or flat sorting, which consists in counting, for a question, the number of answers given for each modality.

The count of missing answers and answers without object is performed during the simple sorting because PRAGMA views these situations as modalities.

These frequencies are calculated and updated at the time of data digitization, then different control operations, coding, recoding and selection. They are edited in real time on the screen window concerning each question. This screen-window presents each question as a count table. Under the question wording, each line isolates a modality and its frequency in different columns.

PRAGMA calculates two types of frequency. On the one hand, “answers”, that is to say the number of time when the modality has been answered, even if an individual answered several modalities for this question or if he answered many times this modality. On the other hand, “individuals”, that is to say the number of individual, who answered this modality. The answers calculation is privileged by the questionnaires processing, whereas the individuals calculation is of general use within data analysis. By default, PRAGMA displays the frequency per individual.

This direct read-out of results allows anytime and any user to refer to the results of simple sorting. Thus, an organism that enters data concerning its public can easily refer to its results before its data be gathered for a global shared diagnosis. Each user can know data of which he is in charge. A global assessment report is automatically provided during the gathering of data. We can give details about these global results per territorial zones with a little organization. These two calculations are different as soon as the question allows many answers and it is possible to answer several times the same modality.

The global quantitative assessment report exports the « unrefined » results of the simple sorting for each question.

## ***5.2. Global quantitative balance***

The global quantitative balance is automatically established by PRAGMA in real time. It can be exported in a file that has a text format and that can be exploited with a word processor or with a spreadsheet like “Excel”.

The balance is composed by all the flat sortings of the questions (or of the selected questions) that constitute the table lines. The first lines indicate the analysis title and the number of statistical individuals.

For each question, the balance indicates in successive lines:

- The question code, row and formulation;
- The frequency of the irrelevant answers if necessary;
- The frequency of no-answers;
- Each modality code, row, formulation and frequency;
- The total number of individuals and the one of answers.

In the columns, for each question or modality are indicated:

- A. The mnemonic code;
- B. The row;
- C. The formulations
- D. The number of individuals that have each modality;
- E. The percentage of individuals that have each modality;
- F. The absolute frequency of each modality;

The relative frequency of each modality.

The formulations describe the questions and modalities meaning. PRAGMA limits the constraints on the formulations so as to avoid the meaning distortions we observe in many analyses throughout we make coding and abbreviations.

PRAGMA is particularly cautious when it makes the counting so as to avoid confusion during the analysis. The surveys examination software usually calculates the frequencies of the answers that are made by the individuals whereas the data analysis software rather counts the individuals that present a modality. The absolute frequencies are different in case of

“multiple” modalities (when a modality can be quoted several times in the answer). The relative frequencies, which are respectively put in relation with the number of answers and with the number of individuals, are different in case of answer absence or of questions with multiple answers.

The number of answers is inferior to the number of individuals when some of them did not answer. It is superior if some individuals answered several times. According to us, the relative frequency has the advantage to be calculated in relation to the same reference for all the questions. Its interpretation is simpler because the reference to the surveyed individuals is clearer and does not vary from a question to another. The frequency that concerns the answers is less explicit and can imply meaning modifications, especially when the multiple question implies the reference to other statistical individuals.

### ***5.3. Cross sorting***

PRAGMA offers a classical function of cross sorting, which performs cross plans by crossing a pivot question with every other question. However, the users generally prefer proceeding per selection. In addition, ANACONDA proposes to extract cross sorting from the contingency table. Thus, the cross sorting theme will have to be studied within the framework of the analysis software integration.

#### **5.3.1. Cross plan**

For each question, the function « Crossing » performs and exports in the form of a text file, the cross of the displayed question, view as a pivot question with every other question (or with selected questions).

This « cross plan » edits the pivot question into columns. The other questions are edited into lines in the questionnaire order.

Three tables are suggested in the form of three columns blocs:

- The absolute frequencies;
- The relative frequencies calculated in respect to the pivot question;
- The relative frequencies calculated in respect of the other question.

#### **5.3.2. Selection**

The users proceed per successive selections. For instance, if they want to cross the type with other questions, they successively select women, then men, from the corresponding modalities. Every time, they export the partial assessment report obtained with the selection

of questions that interest them. Then, they put the partial flat sorting get for women and men together in order to obtain the cross tables.

The interest of this proceeding resides in its simplicity. The time lost during the manipulation is gained in understanding.

The selection also allows comparing complex profiles.

#### ***5.4. Data sheet***

The data sheet is a simple standard for the data analysis.

It is primarily a text file.

All the individuals to analyze constitute the table lines. The characters are the columns table.

The first column table contains the identification code of the individual (four letters or figures in principle). The first table line contains the mnemonic codes of characters.

Each line contains the identification code of the individual following by all the answers of this individual in a Boolean form, that is to say 1 if the individual owns the character corresponding to the column, otherwise 0.

PRAGMA automatically implements and exports the data table from the selected characters and, possibly from the selected individuals.

#### ***5.5. Synthesis question***

The term « synthesis question » refers to several functionalities that we obtain in the same way, by adding a new question, whose answers get inferred from one or many questions. The addition of synthesis questions presumes that the questionnaire grows following processing from the original questions. The questionnaire then contains two kinds of questions: the original questions whose answers are gathered among individuals and the synthesis questions whose answers are inferred. The term « synthesis » refers to the question constitution from many questions. It is necessary to consult the data analysis, in which reference to questions is not direct anymore, and in which characters are separate, in order to understand that we can obtain a synthesis question as well from the development of a question as from the combination of many others.

We mentioned above the possibility to create a synthesis question after each coding, recoding and characters selection operation. This synthesis question would allow showing and memorizing the question progress. It is justified for the coding which often modifies the meaning of the question, which one gives for instance information about the socioprofessional class of the individual and not about its profession anymore. The recoding will more and more reply to different aims of analysis and communication. It will rather consist in a recoding streak, of which it will be important to keep a trace. Likewise, the selection often corresponds to several developments of the question that will be consequently memorized.

Except for these cases, the synthesis question allows solving situations without object, if it is more expedient for the communication to ask the question « How many dependent children do you have? » in two times:

Do you have dependent children?

Yes

No

If yes, how many?

1

2

3 and more

Only one question is more adapted for the processing:

How many dependent children do you have?

None

1

2

3 and more

« None » correspond to the answer « No » for the question « Do you have dependent children? ». 1, 2, 3 and more correspond to modalities for the question « If yes, how many? ». The synthesis question can then be automatically constituted.

Another case settles the matter of questions with multiple answers by splitting them into as many questions with yes or no answers as they involve modalities.

For instance, the question

Among the three following activities, which ones do you practice?

Backpacking

Football

Tennis

Could be split into three questions:

Do you go backpacking?

Yes

No

Do you play football?

Yes

No

Do you play tennis?

Yes

No

Of course, the synthesis question will also help to synthesize several questions into only one question:

For instance, the questions:

What is your current way of life?

Alone, with or without children

A couple, with or without children

Do you have dependent children?

Yes

No

If yes, how many?

1

2

3 and more

Can be synthesized like this:

Kind of household:

Alone without children

Alone with one child

Alone with two children and more

A couple without children

A couple with one child

A couple with two children

Large family

The interest of synthesis questions is that their answers are automatically inferred from gathered data. However, the automation is not integral, it is prepared by a coding or recoding proposition « incorporated » into the software. Then, it is performed in an automatic or transparent way for the software user.

The most difficult operation to automate in this way is the coding. The « automatic » coding of measures can be performed from parameterized thresholds. The text coding cannot be completely automated. We can advice and help with a dictionary soliciting the user for every new answer. The dictionary stores the category for a future automatic determination when the answer happens again.

The recoding, as well as every establishment of a close question from modalities belonging to one or several close questions, can be prepared by showing from which associations or combinations of codes the new modalities are obtained.

The characters selection can also be prepared.

The preparation consists first in deciding the process of coding, recoding, selection or synthesis. This decision is either determined by an expert knowing the indicator, or either memorized after a first experimentation. It can result of a shared experimentation. It is stored as parameters and it is then repeated. The parameters are then modifiable and updatable.

### ***CONCLUSION: PROSPECTS***

Having gathered basic principles of the PRAGMA conception, recommendations and suggestions derived from the experience of its use since more than fifteen years, within different territories in Europe, concerning diversified themes, allow us to establish more precise working perspectives for the implementation of the « CATALYSE toolkit » as well as for the perspective of CATALYSE tools on-lining.

We have already begun from these assignments, the definition of specifications of two technical specifications. They are about:

- A cross-platform version of PRAGMA in JAVA language and
- An online version of PRAGMA in PHP/MySQL

These two versions will be eventually multilingual. The whole developments will be realized with free software. They constitute an ambitious task of which we better control the issues thanks to the experimentation of a « migrants » diagnosis and assessment guide begun at the level of seven observatories in Spain. This guide, resulting from a first homogenization stage at the Spanish level, organizing the synthesis of the « European » guide at the level of CAENTI, has been ended in April 2006. A first version of PRAGMA limited to the data gathering has been suggested the same month in order to begin the implementation and the experimentation of this guide within the framework of five diagnosis in Spain. This experimentation is important because it is simultaneously about the contents understanding, the accessibility of technical solutions and about in situ uses. This experimentation provided us with many learning.

In terms of contents and uses, these learning refer to the reinforcement of the partnership training and leading in order that the whole involved actors share a common language.

At the technical level, which one especially interest us here, the disassociation between the part concerning data-gathering and the part concerning processing allowed reinforcing the accessibility and the sturdiness of PRAGMA in terms of shared data-gathering, which is often performed by users who do not own a computational background, even an office automation background. A version « gathering », rided of functions which help upstream to the questionnaire digitization and downstream to processing, strongly limits risks of takedown mistakes and incorporates less distortion risks The technical follow-up task

of the data-gathering is simplified, the follow-up can concentrate upon the contents understanding.

We also performed a first prototype of PRAGMA, online PRAGMA, also limited to the data entry. It has been recently implemented, in November 2006, with the « migrants » guide. We get two important learning from this experimentation.

Firstly, the training doing is more important with the online version. This version tolerates less disparities. A distinctive mistake can lead to a general dysfunction. Systems of access rights are more complex and more restrictive. Beforehand, we should teach every user before the implementation of the online version to avoid that clumsiness of some users periodically block the whole system and lead to relational difficulties within the partnership.

Secondly, because of disparities at the level of computing equipment and in terms of computational skills, we should consider, at least for a period, the conjoint use of these two PRAGMA versions, personal and online, on a same site. In spite of they are two very different products at the technical level, they must coexist and most of all provide the same database.

Based on these learning, we follow the writing of specifications concerning the data entry module of the cross-platform version and the online version. These tasks will be closely led because both data-entry versions should have similar interfaces, provide the same database and belong to a same rights system. At the level of interfaces, it is imperative that users do not realize the version modification. Even if both versions do not strictly use the same database for technical reasons, it is imperative that conceptual models of data are well articulated and that these versions provide the same database. The rights system, necessarily more complex for an online version, must allow importing data from personal cross-platform versions with the same procedures and the same level of security.

These two versions currently own a module of questionnaire digitization, which one allows storing a « paper » questionnaire and updating the digital questionnaire. We will be surely led up to prefer the online version, which will be eventually more complex insomuch as the digital questionnaire gives more possibilities when it is separate from the « paper » logic. It would be enough to plan import-export interfaces of questionnaire and then to establish the documentary specifications of the questionnaire on the base of the suggestions of this report.

At the level of data processing, we will begin the writing of the technical specifications about a cross-platform version of PRAGMA, which actually contains the three important functions groups: questionnaire digitization, data-entry and processing, which are involved in the processing. Processing constitute a dedicated task more complex to perform from a distance. Our priority consists then in developing a multilingual and cross-platform version of PRAGMA, which one could make for the whole processing and even if data have been entered online or not.

A short-term task, which will allow simplifying processing, is the concrete definition, on the basis of the « migrants » guide and « European » guide of synthesis questions, allowing partly automating the data processing. The aim is to incorporate these synthesis questions within the data entry modules in order that « hands-on » users benefit from more results at their level. This task involves a doing about the definition of processing protocols, question per question, on the basis of five diagnosis led in Spain with the « migrants » guide, as soon as the data gathering will be ended. We will incorporate the experience of their last diagnosis processing by the other actors of the CAENTI in order to incorporate these synthesis questions, with the « European » guide within the PRAGMA data-gathering module, which will constitute the basic tool of the CATALYSE toolkit suggesting immediately a streak of simple results.

The other technical developments are situated within the PRAGMA integration, within the framework of a territorial information system. We also overstepped our aims about the conception of this system as shown in report “Specifications for the PRAGMA integration with the software of qualitative data analysis ANACONDA and NUAGE” (deliverable n° 55).

These developments refer to three tasks:

1. The development of data entry functions more complex than those which are executable with a « paper » questionnaire. This last one is linear and limited whereas a digital questionnaire can contain many questions and will only show the relevant questions according to the user answers. That involves a different logic of skipping over because most of answers dismiss important questions blocs, for instance we will dismiss questions concerning health problems for persons who declare to be in good health. It should be also possible to repeat some questions blocs as often as the respondent specified it, for instance we will repeat questions about dependent children (sex, age, etc.) as often as the respondent said having children. That involves a distinction between the questions list and the questionnaire process

because some questions can be repeated several times. On this basis, we can consider suggesting model questions as the question yes/no that is often used.

2. The precise definition of the whole set of stages concerning the processing protocol at the general level and at the level of the European guide, as main thread and as concrete example. This involves the automation of sorting, calculations and documents output, as well as the programming of processing sequences in order that stages correspond to time of human decisions. These decision-making stages involve either the intervention of specialists or are directly performed by users. That involves delimiting the documents specifications, which allow the preparation of these decisions and which are subject to internal and public publications of results within the framework of the participative governance of development partnerships. Some sequences of this protocol will recover the present functions of PRAGMA, especially the assessment report edition, the coding, the recoding, and the selection of data subgroup. Other processing tasks could be made according to methods of data analysis, such as cross sorting from contingency tables. Other stages should also be integrated within a continuous process, such as cartography.

3. The third task concerning the physical integration of statistical quantitative and qualitative analysis software, PRAGMA on the one hand and on the other hand ANACONDA and NUAGE, once that this two software will have been integrated.

## ***BIBLIOGRAPHY***

FERNANDEZ QUINTANILLA, J., MAHIA CORDERO, J., GIRARDOT, J.-J., MASSELOT, C., 2006: "ACCEM observation strategy", in: *Acts of International Conference of Territorial Intelligence*, ALBA IULIA (Romania), September 20<sup>th</sup>-22<sup>nd</sup>, 2006, 8 p. <URL: <http://www.territorial-intelligence.eu/telechargement/albaiulia2006/Alba06-Fernandez.pdf>>

GIRARDOT, J.-J., 2006: "Activities and prospects of CAENTI". in: *Acts of International Conference of Territorial Intelligence*, ALBA IULIA (Romania), September 20<sup>th</sup>-22<sup>nd</sup>, 2006, 9 p. <URL: <http://www.territorial-intelligence.eu/telechargement/albaiulia2006/Alba06-Girardot1.pdf>>

GIRARDOT, J.-J., 2006: "Activities and prospects of research activities concerning tools of territorial intelligence for sustainable development actors. Work Package 6 "Tools for Actors" of CAENTI", in: *Acts of International Conference of Territorial Intelligence*, ALBA IULIA (Romania), September 20<sup>th</sup>-22<sup>nd</sup>, 2006, 8 p. <URL: <http://www.territorial-intelligence.eu/telechargement/albaiulia2006/Alba06-Girardot3.pdf>>

SANCHEZ, C., GIRARDOT, J.-J., 2006: "Specifications of the contents of the European Guide of Diagnosis and Evaluation.", in: *Acts of International Conference of Territorial Intelligence*, ALBA IULIA (Romania), September 20<sup>th</sup>-22<sup>nd</sup>, 2006, 25 p. <URL: <http://www.territorial-intelligence.eu/telechargement/albaiulia2006/Alba06-Sanchez.pdf>>

GIRARDOT, J.-J., 2005: "Concepts, principes et outils de la méthode CATALYSE", in : *3e colloque international du Réseau Européen d'Intelligence Territoriale*, LIÈGE (Belgique), 20 et 21 octobre 2005, 5 p. <URL:<http://mti.univ-fcomte.fr/reit>>

VAN DER LEEUW, S., FAVORY, F., GIRARDOT, J.-J., 2004: “The archaeological study of environmental degradation. An example from Southern France.”, in CH. L. Redman & al., “*The Archaeology of global change*”, Smithsonian Institution, WASHINGTON, 112-129.

GHIGLIONE, R., MATALON, B., 1978: “Les enquêtes sociologiques. Théorie et pratique”, Armand Colin, PARIS.