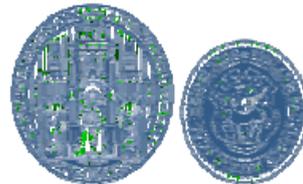MICRODIS

D 4.4.4: Report on lessons learned in common coding and standardized templates for multi-datasets

May 31 2010

Prepared in Collaboration by :

Centre for Research on the Epidemiology of
Disasters (CRED)
Université catholique de Louvain
30.94 Clos Chapelle-aux-Champs
1200 Brussels, Belgium

UniversitätsKlinikum Heidelberg,
Im Neunheimerfeld 342,
D69120 Heidelberg,
Germany

Contents

1. Project background and assessment tool

The MICRODIS project is funded under the by the European Commission under the Sixth Framework Programme. Its overall goal is to strengthen preparedness, mitigation and prevention strategies in order to reduce the health, social and economic impacts of extreme events on communities in Asia and Europe.

The consortium consists of seventeen partners amongst whom there are academic and policy expert institutions from across Europe and Asia who are specialized in key areas of disaster-related health and social science disciplines. In total, MICRODIS Integrated Field Surveys have been conducted in 9 sites in Asia and Europe (see Table 1).

Table 1: Partner institutions and the Survey sites

Partner	Survey site	Thematic working group member
University of Indonesia (UoI)	(Bojonegoro)	Health Working Group (HWG)
Voluntary Health Association of India (VHAI)	Jagatsinghpur	HWG
Hanoi School of Public Health (HSPH):	Hanoi	HWG
University of Northumbria (UoN)	Tewkesbury Morpeth	Social Working Group (SWG)
University of Delhi (UoD)	Bahraich	SWG
Xavier University (XU)	Southern Leyte	SWG
Citizen's Disaster Response Center (CDRC)	Albay	SWG
HUE College of Economics (HCE)	Quang Nam	Economic Working Group (EWG)

In order to assess the health social and economic impacts of floods the MICRODIS consortium has developed a MICRODIS assessment tool (MAT). The MAT was composed of a set of thematic questionnaires which cover questions regarding the health economic and social impacts of disasters. The three thematic working groups (health, economics, social) each developed a core and extended questionnaire. In addition to the thematic questionnaires a MICRODIS common core questionnaire was developed, which included demographic information and main questions needed by each group (see Figure 1).

Each survey team in Asia and Europe put together their country and site specific questionnaires, using the MICRODIS common core and each thematic core questionnaire. In addition, depending on the thematic focus, each country team could further choose questions from one extended.

Figure 1: MICRODIS ASSESSMENT TOOL (MAT) structure

```
┌──────────────────────────────────────────────────────────────┐
│                PART I: MICRODIS CORE Questionnaire              │
└──────────────────────────────────────────────────────────────┘

┌──────────────────────────────────────────────────────────────┐
│ PART II: THEMATIC CORE Questionnaire                          │
│  ┌────────────┐   AND  ┌────────────┐   AND  ┌────────────┐   │
│  │ Health Core│        │ Social Core│        │ Economic   │   │
│  │            │        │            │        │ Core       │   │
│  └────────────┘        └────────────┘        └────────────┘   │
└──────────────────────────────────────────────────────────────┘

┌──────────────────────────────────────────────────────────────┐
│ PART III: THEMATIC EXTENDED Questionnaires                    │
│  ┌────────────┐   OR   ┌────────────┐   OR   ┌────────────┐   │
│  │ Health     │        │ Social     │        │ Economic   │   │
│  │ Extended   │        │ Extended   │        │ Extended   │   │
│  └────────────┘        └────────────┘        └────────────┘   │
└──────────────────────────────────────────────────────────────┘

┌──────────────────────────────────────────────────────────────┐
│ Country adaptations of questions and translation in to local  │
│ language                                                       │
└──────────────────────────────────────────────────────────────┘

┌──────────────────────────────────────────────────────────────┐
│         MICRODIS SIDE SPECIFIC ASSESSMENT TOOLS                │
└──────────────────────────────────────────────────────────────┘
```
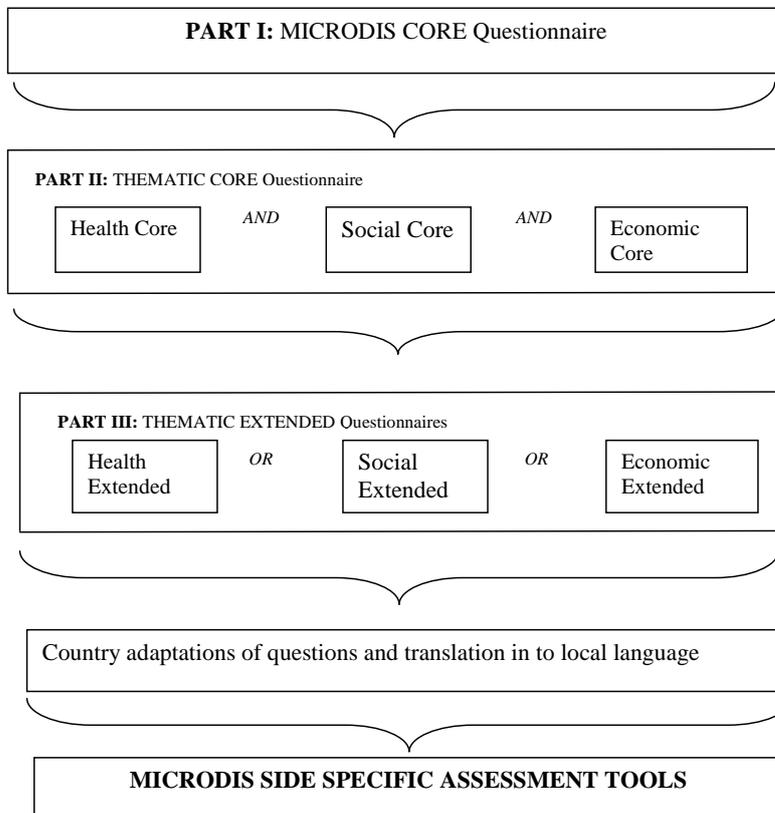
Source: D 5.1.1

2. Integrated Codebook and Data shell

2.1. Rational

Although every team used the MAT slight modifications and local translation of the site specific questionnaire were essential for a successful survey implementation. As a result of these adaptation measures, the straightforward comparison of some questions across sites and countries was therefore markedly challenged.

However, the need to be able to make a cross country, interdisciplinary comparison remained. This made adequate data integration including good data standardization and management strategy were essential. It lead to the need to develop a common data shell in which each of the MICRODIS field data can be recorded in a standardized format using common codes. Further, the creation of a common codebook was necessary to describe each of the variable codes and answer value found in the data shell.

2.2. Objective

The objective of the MICRODIS integrated data-shell (MID) and codebook (MIC) was to improve data comparability across country sites; as well as to facilitate cross country / inter-site data analysis.

2.3. Method

The development process of the MID and MIC were lead by the UCL and UKL-HD and achieved in collaboration with each survey team during the third year of the MICRODIS project. The task started with an intensive work-shop in April 2009 in Brussels with members of the UCL and UKL-HD to finalize the strategy. A broad overview of the strategy can be found in Table 2.

Table 2: Development of the MIC and MID

STEP	Activity	Validation check	Outputs
1	• General review of the MICRODIS common core as well as the health core and extended parts of the used in the nine sides specific MICRODIS questionnaires; • In-depth comparison and analyses of the variable and answer codes from the different questionnaires; • Allocation of standardized variable and answer codes.	Meeting with the members of the health working group in Orissa, India. The objective of the working group meeting was to review the standardized codes and reach a consensus about the coding. Changes were made in accordance with all HWG members.	Core Integrated Codebook; Health Integrated Codebook.
2	• General review of the social questions (core and extended questionnaires); • In-depth comparison and analyses of the variable and answer coded used across the different sites; • Allocation of standardized variable and answer codes.	The UoN reviewed and cross-check all new codes for the social questions. Changes were made in accordance with the coordination team.	Social Integrated Codebook
3	• Review of the economic questions and their coding across the different sites. • In-depth comparison and analyses of the variable and answer coded used across the different sites; • Allocation of standardized variable and answer codes.	The coordination team reviewed and cross-check all new codes for the economic questions. Changes were made in accordance with the coordination team.	Economic Integrated Codebook
4	• Merging the integrated codebooks (Core, Health, Social and Economic) into one document.	Once all the integrated codebooks were merged to one document it was send to every consortium member for comment and review. Changes were made in accordance with the coordination team.	MICRODIS Integrated Codebook (MIC)
5	• Entering data variables and answer codes according to the MIC into SPSS.	Once all the data and answer variables were entered in to the SPSS, the data shell was shared with every consortium partner for pilot and review. Changes were made in accordance with the coordination team.	MICRODIS Integrated Data shell (MID)

A more detailed description of the steps undertaken in order to develop MICRODIS integrated codebook (MIC) and data shells (MID) can be found in Deliverable 5.1.1 *"Report on common dataset template, data entry shell and validation checks"*. The document is available on the website: www.microdis-eu.be.

3.   Challenges faced and lessons learned

3.1 Differences of Survey times and time lapsed after disasters

Since most of the surveys were conducted at different time intervals after the disaster occurrence, it was necessary to consider the recall period particularly for disease information (see Table 3). A few of the surveys were conducted between a few days after the disasters to some which were conducted 2 years after the disaster.

This posed a specific challenge to the time dependent questions specially related to health. This was solved in the final coding by standardizing the definitions of before, during and after the disaster and comparing only the relevant information across sites.

Table 3: Time interval between the disaster studied and survey conducted

No	Site	Disaster studied	Time of Disaster studied (mth, yr)	Time of survey (mth yr)	Thematic area
1	Albay, The Philippines	Windstorm/ flood, volcanic eruption	Sept- Nov 2006	Dec 2008	Social
2	Bahraich, India	Flood	July - Sept 2007	Oct 2008	Social
3	Bojonegoro, Indonesia	Flood	Dec 2007	Nov 2008	Health
4	Hanoi, Viet Nam	Flood	Oct-Nov 2008	Feb-Mar 2009	Health
5	Morpeth, UK	Flood	Sept 2008	Aug- Nov 2009	Social
6	Jagatsingpur, India	Flood	Sept 2008	Nov 2008	Health
7	Quangnam, Viet Nam	Flood	Oct -Nov 2007	July - Aug 2009	Economic
8	Southern Leyte, The Philippines	Windstorm, Earthquake	Feb 2006	Dec 2008	Social
9	Tewkesbury, UK	Flood	July 2007	Jan 2009	Social

3.2 Order of questions altered

As described in the sections above, the structure of the questionnaire was set as the MICRODIS core which included questions with regards to demographic information common to all sites and all groups. The thematic core contained questions which were to be included at all sites and by all thematic groups as well.

However, the extended thematic questionnaires contained questions specific for each of the thematic groups followed by the country specific questions. Each country team assembled the questionnaires to suit their site. This of course meant that for sake of continuity in their respective questionnaires, each team changed the sequence of appearance of the questions in their individual questionnaires. Since the questions were not coded before hand, this meant every team had the same question appear in different sections of the questionnaire and also with a different code.

Each team followed their own sequence of questionnaire coding and we were faced with the challenge to re-code the questions retrospectively to standardize the question codes. This was a problem because each team had collected it in a different section altogether either and some of them had already entered the data. If the questions were standard it was still feasible but there were other challenges with the questions themselves and these are discussed below.

3.3 "Core" questions merged with "extended" questions

The MICRODIS core questionnaire and the thematic core questionnaire were essentially the questions that were to be used for cross country comparisons. However, since the questionnaires were reformulated and question sequences changed the questions in the core and the extended sections were merged by some of the teams. This made it a challenge to revisit each of the questionnaires retrospectively and to identify and separate them into core and extended again. This was necessary to bring together the core and thematic core again and to have the right sequential coding across the questionnaires and hence the common data shell and the codebook.

3.4 Time lines of the questions were adapted

The standard questionnaire captured information before, during and after the disaster. However, since the time period before and after was open to interpretation, some country teams modified their questions to include words such as one month after the disaster, or just after the disaster! This meant that the time lines for the information captured differed between the teams. In the final codebook however, the 'after' data was grouped and we arrived at the final before, during and after the disaster codes to standardize the data for comparison across the sites.

3.5 All timelines for questions not captured at all sites

The main idea with the before, during and after questions was to collect information regarding how things changed from baseline. The standard questionnaire captured relevant information for the time before the disaster, during the disaster and after the disaster. However, some of the teams asked questions only for status before the disaster and during the disaster period.

Others captured the information for before and after ignoring the "during" period. In some before, just after and at present (time of interview) which was a challenge to interpret and standardize.

In the example shown in Table 4 below, some teams added details such as "displacement" after the disaster. In such cases we revisited every question in every questionnaire with the time line before, during and after period and combined the responses to arrive at a standard before, during and after disaster information to allow comparisons. This was done keeping in mind what was comparable and what was not as the data was already collected with a certain definition preset by the teams.

Table 4: Timelines not captured at all sites

Example: Different time lines of data capturing

☐ What is your primary source of drinking water?
      1. Answer the question for before the disaster
      2. After the disaster if not displaced
      3. After the disaster if displaced
      4. Now
 ☐ Other teams asked the same question, before or after (without interest in the displacement factor)

Solution in common data shell and codebook
☐ Combine 2*3 as the after timeline response irrespective of the displacement.

3.6 Rephrased questions

Given that English was not the first language at a majority of our study sites the teams rephrased some of the questions to suit the local contexts. This meant that the rephrased questions had to be revisited and checked if the same meaning was preserved. This also posed a great challenge for questionnaire coding because the questions were no longer in the same verbatim. This was also attributed to the translation to and back translation from English and the individual local languages.

3.7 Unclear definitions of grading of responses

Not only were there differences in recording the impact of disaster before, during and after as mentioned above, but the grading of the responses were also altered in some cases. For example in Table 5 below, the option 3 was not recorded by one of the teams as they judged "most of the time" and "a good bit of time" to be similar and merged the categories. Additionally, it was also unclear to the interviewee what the interviewer meant by these 2 closely related terms.

This made interpretation difficult and since the data was already collected we faced the challenge of grouping data in a way that we were comparing the right information across the sites. As a solution in the common data shell we combined option 2 and 3 to capture all data

and to bring in comparability as the teams that had collected the data in the combined form could no longer separate it into 2 separate categories. But the teams who collected the data in separate categories could easily combine them.

Table 5: Unclear definition and grading of responses

Example: Unclear definitions of answer codes
☐ **Have you felt calm and peaceful?**
1. All of the time
2. Most of the time
3. A good bit of the time *(not recorded!)*
4. Some of the time
5. A little of the time
6. None of the time
Solution in common data shell and codebook
Options 2 and 3 combined

3.8 Answer codes options "added"

The answer codes were agreed and pre-decided. However, when the country teams conducted their pilot studies some additional responses emerged as major players. These responses that did not feature in our original circulated questionnaire were then added as additional response codes. This meant that at the end of the survey when we compared all sites we had additional responses collected. These had to be included in the common database. Therefore each questionnaire was scrutinized and the additionally "added" answer codes were collected across sites and added so that all teams had their answer codes feature in the common data shell and codebook. Below in Table 6 you can see that in the original questionnaire we had 12 answer options for Relationship to head of household. However, in the Asian context where large families live together with in-laws, additional relations were to be added to include all the members living together in that particular household.

Table 6: Additional Answer codes added

Example Answer codes added
☐ Relationship to head of household
■ 12 answer codes in original questionnaire
■ 13+ in some questionnaires
☐ Highest level of education
■ 12 answer codes in original questionnaire
■ 13+ in some questionnaires
Solution in common data shell and codebook
☐ 13+ the 95-99 options so 18 answer codes in the final codebook.

3.9 Multiple responses offered for the same question

The answer options for most questions were either/or responses in the standard questionnaire. However, some teams modified the answer codes to allow the respondents to give more than one answer response to a given question. This meant that recording these responses was essential. Therefore in the final data entry and coding we had to separate the answer options into independent questions with the responses as Yes and No (see table 7 below). This allowed us to capture all information collected. But this also meant that the final number of questions increased and thus the coding became more challenging. Besides, not every team offered multiple answers options.

Table: 7 More than one response offered

Example 1 : 2 or more responses possible for single question

☐ Were houses, yards or roads surrounding your house flooded?
    ■ 1. Road was flooded
    ■ 2. House was flooded
    ■ 3. Yard was flooded

Solution in common data shell and codebook

☐ 3 Separate questions with Y/N answers
    Was the Road surrounding your house flooded Y/N
    Was your House flooded Y/N
    Was the yard surrounding your house flooded Y/N

3.10 Answer categories Adapted – either merged or separated

In some of the questions, the teams merged or separated the answer categories depending on individual relevance. This meant that a certain site had collected data for a particular category separately while the others had merged it into one. This meant that if we did not want to lose any information, we had to carefully analyse every questionnaire across the sites and find the merged / separated answer codes. An example with regards to the source of the water is given in Table 8 with the final solution we came up with.

Secondly, as seen in the example the sequence of the answer responses (1,2,3) also changed for the final coding and thus every team had to recode the data completely including the answer codes for each category to achieve standardization. For example option 9 in original questionnaire was surface water and in the final codebook it was option 10.

Table: 8: Answer categories merged or separated

Original questionnaire	One of the questionnaires	Final code book
1- tap in house	1. Piped water/PDAM into dwelling	1. Piped water/PDAM into dwelling
2- Communal tap	2. Piped water/PDAM into yard	2. Piped water/PDAM into yard
3- Tubewell/borehole	3. Public tap/hydrant	3. Public tap/hydrant
4- Protected dug well	4. Bored well (manual pump, engine pump)	4. Bored well (manual pump, engine pump)
5- Unprotected dug well	5. Covered/ Protected dug well	5. Covered/ Protected dug well
6- Water from protected spring	6. Open/ unprotected dug well	6. Open/ unprotected dug well
7- Water from unprotected spring	7. Covered/Protected spring	7. Covered/Protected spring
8- rain water collection	8. Open/ unprotected spring	8. Open/ unprotected spring
9- surface water	9. Rain water	9. Rain water
10- Tanker truck	10. Surface water (river, lake, pool, dam, canal stream, irrigation channel)	10. Surface water (river, lake, pond, pool, dam, canal stream, irrigation channel)
11- mineral water	11. Water tanker truck /water container/water seller	11. Water tanker truck /water container/water seller
88- OTHERS:, SPECIFY _____	12. Ask to neighbour who was not flooded	12. Ask to neighbour who was not flooded
	13. Bottled water for one time use	13. Bottled water for one time use
	14. Refill bottled water Others….	Others……

3.11 Standard codes for general responses

The codes such and Yes and No were not uniformly coded as 1 and 2. In some cases option 1 was "Yes" and in others 2 was "Yes". It was the same for the "others" response category. Some teams coded it as 88, some as 99 etc. Similarly there was no standard coding for if the information was not available, was not recorded or was not known to the respondent. Given that the final codebook contained all the questions for all the thematic groups and all the countries we needed a standard code for these options because for example the social group did not record the health extended questionnaire and vice-versa. Thus all the teams finally coded the yes, no, others etc. options as shown in Table 9 below.

Table 9: Standard general codes

Code	Value Label
1	Yes
2	No
95	Not recorded
96	Other
97	Not applicable
98	Don't know/forgot
99	Missing

4. Achievements

4.1 Successful development of the Health and Core Integrated Codebooks

Since both, the UCL and UKL-HD were members of the Health Working Group (HWG) it was decided to start the recoding process with the health questions. It was also agreed to review the questions from the MICRODIS Common Core questionnaire as it was used by every country team. The first draft of the recoded codebook was achieved during the one week working meeting in Brussels in April 2009. The first quality was then performed successfully by revisiting the codes separately by UCL and UKL-HD. A third neutral check was than conducted by the UCL statistician (Jose Rodriguez). The first draft of the *Health Integrated Codebook* and *Core Integrated Codebook* were shared with the all HWG members in the first instance for review.

In addition, both documents created an excellent working basis for the planned meeting with the members of the HWG in India. The HWG meeting was planned as a validity check during which every single health question was reviewed in order to create a better understanding of the differences among questionnaires and put emphasis on the need for common coding.

A common agreement in that each partner will adjust their core questionnaires (data set in most cases) to a common standard was accepted during meeting. Having also reached a successful agreement of the health codes, the review process went on to the social and economic questions of the MAT.

4.2 Successful development of the Social Integrated Codebooks

The strategy of common codification was extrapolated to the Social Working Group (SWG) with the consensus of the SWG leader. An individual question review for the social questions was successfully conducted and the *Social Integrated* was created. Once the codebook was created a member of the UCL met with SWG member (UoN) in Brussels to clarify the coding procedure. A separate validity check was than performed by the UoN.

### 4.3 Successful development of the Economic Integrated Codebooks

The HUE team was the only team with an economic research focus and thus they created the common codes for the economic questions. The *Economic Integrated Codebook* was then sent to the UCL team for quality control. A second review of every economic question was undertaken to ensure that every team would be able to apply the codes suggested. In agreement with the UKL-HD and HUE team necessary modification were made by UCL.

4.4 Successful development of the MICRODIS Integrated Codebook

Once all the thematic codebooks had been finalized they were merged together to the *MICRODIS Integrated Codebook* (MIC*). With the MIC* an instruction booklet on how to code and enter the data into the database in order to minimize entry errors was created.

Validity checks were conducted by each of the three thematic working groups for the health, social, economic, and the core questionnaires - to ensure that all the question codes and answer codes in the final data shell and codebook included questions and answer values from each site-specific MICRODIS questionnaire. Once the comments were received UCL (in consultation and agreement with UKL-HD) the necessary modifications were made. The final MIC was circulated for review to all MICRODIS partners in September 2009. A brief outline of the different component of the MIC and layout are shown in Figure 2 and Figure 3 below.

Figure2: Components of the MIC
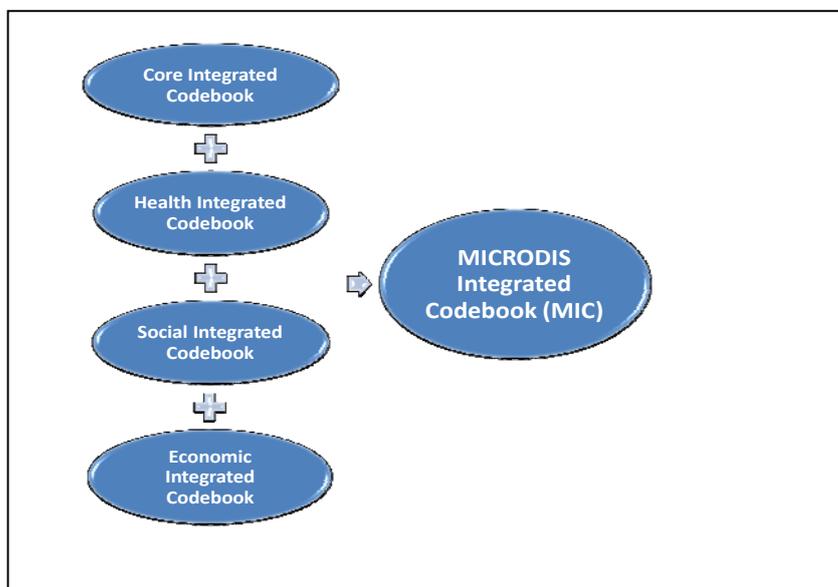
For all four sections (Core, Health, Social and Economic) the data was separated by household and individual level data in codebook as well as in the data shell. The MIC has the following sections (see Figure 3).

Figure 3: Contents of the MIC

4.5 Successful development of the MICRODIS Integrated Data shell

Once the MIC was established the MID was created exactly according to the codebook. The MID together with MIC gave partners the basis to be able to recode their datasets according to carefully chosen codes which can be standardized across all country sites.

4.6 Successful implementation of MIC and MID

Once the MIC and the MID were declared as final drafts they were given to every consortium member and the recoding process began. Great effort was made by every Survey team to recode the dataset. Although this was a time consuming task each team successfully completed the task. All recoded datasets were submitted to the coordinator by the beginning of January 2010.

4.7 Improvement of knowledge and understanding of differences among questionnaire data

As every survey team was granted the permission to adapt the MAT according to their social and cultural background no questionnaire was longer the same. Adaptations were done in the wording of the questions, adding new questions, adjusting the answer possibilities etc. This evidently lead to the situation where no survey team any longer had good understanding of the other questionnaires – unless the understanding was actively thought individually.

For the purpose of the integrated project it, however, remained important that the survey team as a whole developed an understanding of the differences in the surveys to be able to conduct cross country analysis in the future. Although, the MIC and MID were produced in order to standardize the datasets that alone would not prevent future misunderstandings.

During the recoding process active interaction of the survey teams was promoted. This, for instance, was achieved by a very fruitful working group meeting with members of the Health

Working Group was held in Orissa, India from the 20th to the 23rd of May 2003. In total eight health working group members and one member of the Social working group were present at the meeting. A member of the social working group was invited to the meeting in order to ensure his understanding of the recoding procedure and apply his knowledge to social questions of the MAT. Further the SWG member attended the meeting in order to address issues as how social data can be integrated into the health data later in the data analysis.

Although there were no extended working group meetings for the social and economic working groups' good communication was held up by phone and e-mail contact.

The MIC and MID at Multi Meeting were presented and explained again during the Multi Meeting Workshop (MMW) held in Jakarta in October 2009. The consortium members were given the opportunity to ask questions and express concerns. A discussion session also enhanced again the understanding of the MIC and MID.

4.8 Problems were solved

After all the recoded datasets were obtained UCL undertook a quality check by going through every dataset, variable by variable and double checked whether the variables corresponded to the comment codebook. Every team received a feedback from UCL with the required individual modifications. Further, if a problem was identified by any individual team in the codebook/data shell then the necessary modifications were made and/ or granted to the teams.

In case the country came across any difficulties during the recoding procedure immediate assistance was given by the task leaders to ensure that only changes were made to the MIC and MID if absolutely necessary. The coding errors or difficulties were addressed and solved as quickly as possible and communicated any changes to the MICRODIS consortium.

5   Recommendations based on the Lessons learned

The main recommendations based on the learned can be summarized in short as follows:
- Once questionnaires are fianalized:
  - The language of the questions should be locked to change.
  - Provide the teams with very clear definitions of the time frames - before, after, just after etc.
  - When construction inter-site and inter-thematic questionnaires the codes should be preset and finalized before piloting of the questionnaire.
  - The response options offered should be preset and locked to change before circulating the final questionnaires for fieldwork.
  - In large inter-country surveys, it is pragmatic to keep individual country databases separate for analysis in order to use all the information collected.
  - The information that can be compared should be pre-identified and collected in a standardized way to avoid some of the problems described above.

- Some of the information collected may not be used. Therefore, it is essential to collect data that can be analysed. In addition, detail information, which does not change the overall picture, should be left out.
- During the sampling and questionnaire designing phases, the expected statistical challenges and problems with data merging should be explicitly discussed.
- The expected statistical challenges with data analysis should also be discussed in detail before merging datasets.
- A standard statistical package should be pre-agreed by the participants and staff trained before the data entry and analysis phase of the project.