Quantitative indicators for the measuring progress in the area of Roma Inclusion.
Possible approaches to the issue of ethnic sensitive data collection

Draft report on the initial findings and ideas, September 2011

Background

Since the beginning of the Decade of Roma Inclusion, significant progress has been made at least in one area – governments, international organizations and civil society activists endorse the idea that sound and reliable monitoring is a necessary precondition for making progress in Roma inclusion. There is almost unanimous consensus that reliable data and relevant indicators are needed both to identify priority areas of intervention and allocate resources as well as to monitor the real change on the ground (if any). But that's about all in regards consensus. A major problem – agreeing on some basic demographics of Roma population – is still unmet challenge.

The challenge is not just of research nature. Appearing in the denominator of any indicator, demographic characteristics like population numbers have direct policy and political implications. At the same time, the *EC Communication on an EU Framework for National Roma Integration Strategies*[1] issued in April 2011 calls on all Member States to "ensure that national, regional and local integration policies focus on Roma in a clear and specific way, and address the needs of Roma with explicit measures to prevent and compensate for the disadvantages they face." For that matter the EC proposes that national Roma inclusion strategies are designed with **targeted** actions and **sufficient** funding to deliver them. Both 'targeting' and 'sufficient' entails determining who are the populations that need to be targeted, where are they exactly, what is their number and socioeconomic characteristics, so that their measures envisaged to address the challenges (and resources allocated for the purpose) can safely claim to be sufficient.

This is why the pilot project has explicit component on brainstorming possible ideas and suggesting feasible approaches to the 'quantification' issue.

Data and indicators – the case of Roma

General principle
The research done in the context of the project (and in the wake of the Regional Survey) confirmed earlier findings and the initial hypothesis that it is neither possible nor reasonable to invent and implement "Roma indicators". The very logic of the Decade of Roma Inclusion and other inclusion-targeted initiatives is to integrate Roma in mainstream society. This is also the envisaged outcome of the EU-wide efforts in general and the requested national Strategies in particular. This suggests that for monitoring the progress in regards Roma inclusion (and the implementation of the National strategies once they are elaborated/adopted), standard sets of socio-economic and human development indicators should be applied. This would provide room for comparability with the results of other data collection and monitoring exercises conducted on regular basis by NSI, sectoral agencies and line ministries.

In this regard the challenge is not so much in developing new indicators but

(1) in identifying the universe under study (answering the question "who is Roma") and

(2) feeding the standard indicators with ethnically disaggregated data.

Given the experience so far, including of EU member states, "counting Roma" is not an option. Even assuming it is technically feasible, such an exercise would have produced numerous negative consequences. Even if politically feasible (which it is not), the task would be impossible to solve methodologically due to the vague

---

[1] Communication from the European Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, "An EU Framework for National Roma Integration Strategies up to 2020**",** Brussels, 5 April 2011, COM(2011) 173 final.

definition of "a Roma". This is why a 'second best' approach is suggested based on the idea of **attaching ethnic markers to standard socioeconomic data**. It means that for monitoring the Decade and its priorities standard indicators are suggested but disaggregated by ethnicity using different types of ethnic markers.

Types of indicators

The experience from reviewing various policy documents in the countries participating in the Decade of Roma Inclusion suggests that strict distinction between "input", "output", and "outcome"/"impact" (outcomes and impacts can be difficult to distinguish) indicators should be followed. This distinction is not that clear for many practitioners (including people involved in the EU Roma pilot projects implementation). Even worse, it is not being consistently being made by decision-makers at various levels of government and international institutions with a stake in Roma inclusion.

The experience from the pilot implementation and the related discussions with various stake-holders suggest that clear distinction should be made (and deliberately promoted) between 'National strategies'(that the Member States are requested to prepare or revise by the end of 2011) and the Decade of Roma Inclusion National Action Plans (that all countries members of the Decade have and implement). The National strategies indicators should be falling into the scope of long-term outcomes/impact and those of the NAP should be more operational and short-term. In any case deliberate efforts need to be taken to avoid confusion between different indicators and reporting frameworks (which is already a fact to certain extent).

In regards to National Action Plans, it seems appropriate to use two levels of 'outcome' indicators – one for the level of "goals" of the National action plans (the achieved goal within the given priority) and a 'lower level outcome' for individual "tasks" (or "targets" as they are defined in many NAPs). The level of "actions" require output indicators although in many cases the distinction between outcomes and outputs would be vague due to the way the specific "action" is formulated. Input indicators have operational nature and should be defined by the respective agencies drafting the particular operational plans and activities that are supposed to contribute to achieving the outcome.

It seems appropriate to develop a brief guideline on the construction and possible use of different indicators for the purposes of the National Strategies monitoring and evaluation. Introducing (or suggesting) certain common approaches, definitions and methodologies for collecting the data necessary for the respective indicators will help avoid unnecessary confusion and will simplify the evaluation process later on.

Data sources

For adequate monitoring a combination of data sources is required. It includes (1) "hard statistics" – statistical monitoring data and data from administrative registries, (2) survey data (both regular surveys conducted by National Statistical Institutes (NSI) and surveys conducted by NGOs) and (3) data from community-based monitoring. All three sources provide specific inputs and are supplementary to each other. Hence the major sources of data that can be used for the purposes of progress monitoring are:

- Census data

- Sample surveys (conducted on regular basis by statistical offices like Household Budget Survey, Labor Force Survey as well as custom surveys like the one conducted for the purposes of the pilot)

- Administrative registries

- Line ministries registries (in particular, Ministry of education, Ministry of health)

- Specialized agencies registries (Health insurance institute, National social insurance institute)

- Anonymous surveys conducted on the spot by service providers (labour offices, hospitals)

- Community-based data collection

All these sources provide different type of information which will be used both for ethnic markers and data disaggregation by ethnicity. Different types of indicators described above also require data from different sources. For example

- "impact" can be monitored only in the long run and data from population census is appropriate for that purpose;

- "outcome" has shorter time-frame and data from HBS, LFS and other similar instruments are appropriate.

- "output" and "input" indicators belong to the operational planning of the process and should be monitored on the basis of data from individual institutions reporting systems. Defining the lines of reporting needs deliberate efforts and methodological support for the respective agencies may be necessary. This goes beyond the scope of the pilot however.

In most cases the data sources have the information necessary to feed in relevant indicators. It means that the challenges in most cases are of administrative and not of methodological nature. In some cases however data is not collected on regular basis and specific recommendations are made in this regard in a separate annex.

In some cases data from several sources can be provided for the same indicator. For example, unemployment rate can be calculated and monitored on the basis of census data, on LFS data and on Labor offices registries. In such cases it is suggested that all three are used because they provide information on different aspects of the complex reality. Metadata is also crucial for preventing inaccuracies, misinterpretation or misuse of data.

Possible ways of how ethnically disaggregated data can be generated

Important caveat is necessary here: any quantification of a universe as diverse as Roma will inevitably be an approximation. Different approaches are supposed to provide more reliable estimates of the population in question (and thus of the magnitude of the challenge of inclusion). Any effort needs to be made to make those estimates more reliable than the 'guesstimates' currently floating around – but the former would still remain an estimate. Moving from less reliable to more reliable estimate is however feasible task worth taking.

With this caveat the following options for generating ethnically disaggregated data have been identified:

(1) Disaggregating hard statistics using different ethnic markers (personal identification numbers and territorial tags from qualitative surveys)

(2) Extending the samples of regular sample surveys with Roma boosters

(3) Custom "on the spot" surveys among recipients of social services

(4) Community-based collection of data conducted by data collectors from the communities monitored

All of the approaches above are mutually reinforcing and complement each other and should be seen as integral pillars of comprehensive system of ethnically-sensitive data collection and monitoring. All five can work in the case of Bulgaria and are consistent with Bulgarian legislation and data collection standards. In some cases though additional legislation needs to be developed and passed to ensure full respect to privacy and individual data integrity.

## Disaggregating hard statistics using different ethnic markers

The first approach of the two in this group uses Personal Identification Number (PIN) as ethnic marker. It is based on the fact that census records ethnic affiliation (mother tongue) and individual respondents' unique personal identification number. Administrative and other registries do not maintain data on ethnicity – and so does the PIN (one can determine the sex but not the ethnicity from a PIN). Ethnicity however is registered during census and so (generally) is the PIN. Most of administrative registries use PIN as well. So using PIN as common link between ethnic attributes from census and different data sets, various administrative registries can de disaggregated by ethnicity and ethnic-sensitive indicators can be computed. The diagram below gives an idea of the method:
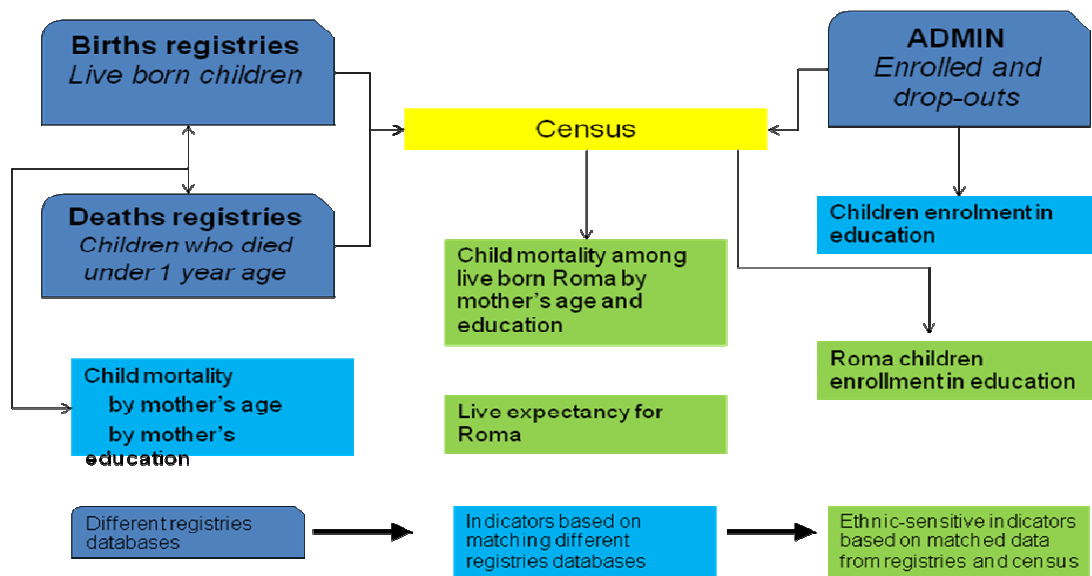
This should be done on aggregate level (not revealing individual ethnic identity.) Matching the census identity with PIN registration in administrative data bases makes possible identification of the representatives from the respective ethnic group out of the total universe of the respective data base. Below are the results of pilot testing of this approach using data from Bulgaria:
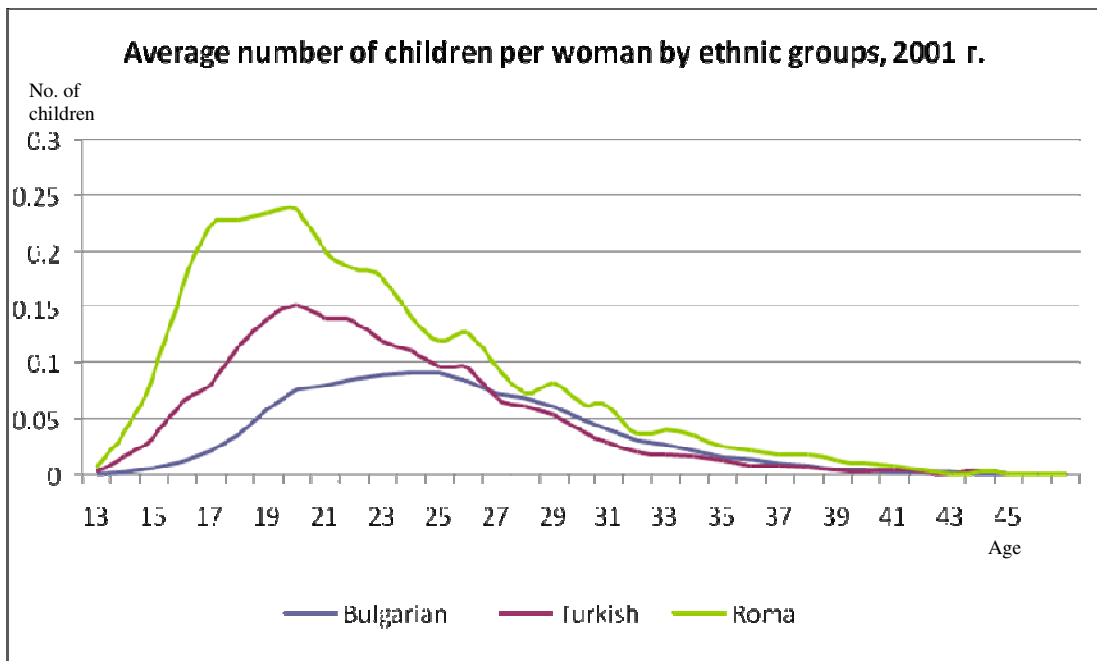
Life expectancy at birth, 2001 Census
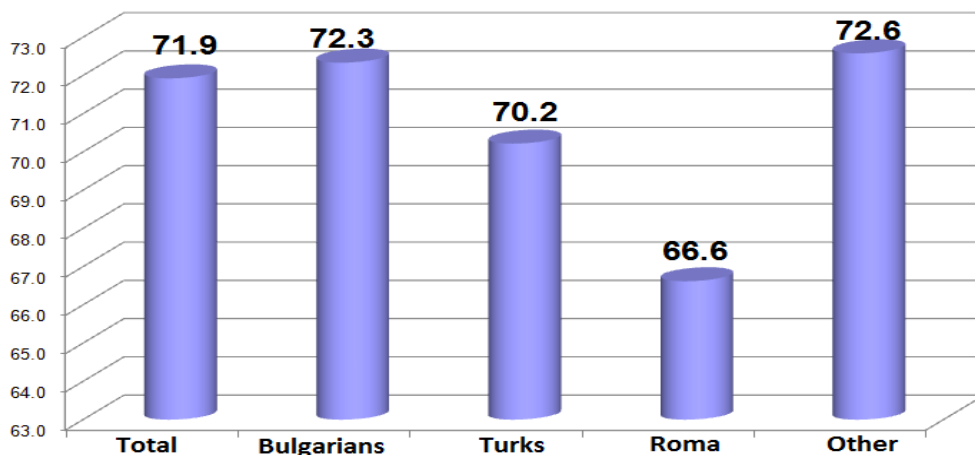
Ethnic group	1992	2001
Average number of children per woman		
Bulgarian	1,41	1,16
Turkish	1,92	1,64
Roma	2,93	2,77
Early (juvenile) birth rate (births per 1000 of age below 18)		
Bulgarian	66,2	41,3
Turkish	283,1	179,6
Roma	690,3	508,8
Extremely young birth rate (births per 1000 of age below 15)		
Bulgarian	3,1	2,4
Turkish	20,3	21,5
Roma	70,1	35,6
Child mortality by ethnic group (deaths per 1000)		
Bulgarian		9.9
Turkish		17
Roma		28

This approach is applicable for extracting national-level ethnically disaggregated data on administrative (incl. population) statistics, registered unemployment, health treatment (both hospitalization and personal doctors visits), social insurance coverage (incl. labor contracts). Indicators like registered unemployment rates, morbidity rates, mortality rates social assistance coverage, formal/informal employment rates can be computed with high level of accuracy.[2] For that purpose however explicit procedures for data anonymization and relevant administrative structure should be in place. Establishing such a structure is one of the proposals of the pilot. This approach could be particularly helpful for generating data and computing indicators in areas that are difficult to address through sample surveys (like health indicators, which are notoriously weak in NAPs and are being monitored primarily through input-level indicators). Examples of outcome-level health indicators that can be obtained using PIN approach include:

- Prenatal, neonatal and postnatal mortality

---

[2] The fact that census data underestimates the number of Roma population is not a problem because the similar degree of underreporting will appear both in the nominator and the denominator. In addition indicators computed on the basis of PIN-tagging can be correlated with other data to improve their robustness.

- Number of not hospitalized births out of the total number of births

- Child mortality by mothers' age

- Roma morbidity (most common illnesses)

- Percentage of Roma with health insurance

- Percentage of Roma covered by screening surveys

- Number of Roma who passed a regular medical check-up

- Number of Roma registered in the system of social service's primary health care

Territorial markers tagging

This approach is based on the fact that most of the vulnerable Roma – those who are the explicit target group of the Decade and EU-level inclusion policies – live excluded also territorially, in separate (often segregated) communities. Thus territorial mapping of those communities is possible.[3] Once a detailed map of Roma-dominated communities is available, ethnic tags based on individual's address can be applied with the assumption that an individual living in an area identified as "predominantly Roma" is Roma. These tags can be used in line ministries registries (particularly Ministry of Education), personal doctors data bases etc. Other indicators calculated based on this approach include:

- Types of dwellings

- Size of the dwelling; m2 per household member

- Average number of members per household

- Average number of households per dwelling

- Child mortality under 1

- Frequency of mother mortality by age and by main death causes

- Frequency of hereditary diseases

- Frequency of sexually transmitted diseases

- Percentage of the children under school age covered by health services

- Percentage of family/mothers who renounce to have basic health cares for their children

- Progress/regress in school desegregation

It should be made clear that these indicators are "indicators for a population living in certain area with certain parameters" and they are not directly "indicators for this or that ethnic group". In cases when the group is highly concentrated (like Roma populations living in segregated settlements), there is a match. But generalizations for 'entire Roma population' are neither possible nor correct. On the other hand, this approach provides opportunities for observing the status (and measuring the progress/regress) in territorial areas with high concentration of Roma, which is the policy priority of Roma inclusion efforts.[4]

---

[3] Similar mapping of Roma communities was conducted in Slovakia and the exercise was successful. UNDP is currently working on the update of the Atlas that is supposed to be completed in early 2012.

[4] 'Roma inclusion' is an intellectual shortcut. When we use this term, we mean 'including the excluded Roma' and not 'including those who are already included'. This is why focusing the data collection and monitoring efforts on the segregated (who happen to be those excluded) is acceptable from policy perspective, although may be insufficient from research point of view.

Ethically-disaggregated data based on territorial markers tagging can be particularly important for monitoring Decade progress in education. One of the objectives of the Decade in this priority area is schools desegregation. If number of children attending a school in Roma neighborhood (presumably segregated) is monitored, the only "indicator" for progress in desegregation will be "decrease of segregated schools enrollment rate". However it won't show where are those not attending a segregated school – attending a mainstream one or not attending school at all. Tagging individual students' address with the ethnic mapping data will provide information on their ethnic affiliation regardless what type of school they are attending. The same approach is applicable to any data using individual's address (like for example medical establishments, employment/unemployment etc.). Other areas in which territorial tagging can be particularly useful include:

Territorial markers tagging is thus complementary to PIN-tagging. But it has some benefits that the latter does not have. To certain extent it can be more reliable because solves the problem with understating ethnic identity during censuses (both due to unwillingness of the respondents to disclose it or unwillingness of the current political elite to include questions on ethnic identity). It is also less susceptible to fluctuations due to changes in political environment (revealing ethnic identity is heavily influenced by the political climate, rise and influence of extremist parties for example). Intensified migration flows, which in the case of Roma community are affecting entire communities, may distort this argument however. Those benefits however come at a cost – it grasps the marginalized, visually excluded segment of the Roma population whilst the probability that integrated Roma will fall out of the scope of the data collection exercise is high.

In any case however using territorial markers tagging is important (and to certain extent – the only reliable) approach that can provide acceptably relevant estimate of the absolute number of the population in question (and not just shares as poverty rates and unemployment rates). The absolute number is crucial for needs assessment and hence for defining numeric targets. If targets (and resources) are determined on the basis of census data, the real needs will be inevitably underestimated.

Roma boosters in sample surveys
Apparently this is the easiest way to get ethnically disaggregated data. Increasing the samples regular statistical data collection surveys (like HBS and LFS) would provide comprehensive information on income, expenditures consumption patterns, employment status and qualification of the labor force. Less data would be available on educational aspects of children and youth (not part of the LF). Still, this data would be important input for monitoring progress under Priority 3 (housing) and Priority 4 (employment).

In reality however constructing the random sample boosters may be a problem, mostly because of the unclear number of Roma population and the "flexible" nature of the very definition of "Roma". One possible compromise is accepting self-identification principle (during census) and constructing a random sample based on the population self-identified as Roma or having declared Roma as mother tongue (ideally both). In this case a Roma booster would bear the "genetic" features (and problems) of the PIN-based methods for statistical data disaggregation and shares both its benefits and detriments. A more appropriate option is constructing a sample using a combination of territorial distribution of the population as reflected in the census, external identification of areas inhabited by Roma population. At the last selection stage however external identification should be complemented by some elements of self-identification (of 'implicit confirmation of Roma identity' as was the case in the representative survey conducted for the pilot project). However this approach is acceptable for random sample surveys and not for panel surveys, which appear as interesting opportunity to generate quantitative data on the status of Roma population in timely and cost-efficient manner (provided that reliable sampling frame is possible to be built).

Custom surveys among social services recipients
This approach entails anonymous questionnaires (usually brief, consisting of just few questions) filled in by recipients of social services on voluntary basis. For example, unemployed person registering at the labor office is invited to fill in a questionnaire in addition to the regular forms. The questionnaire may include field "ethnicity" and is dropped in a sealed box to make linking of the questionnaire with the standard application impossible.

Such approach can be a good source of information both on the ethnic profile of the recipients of social services and of the way their providers work (for example, are there any ethnic-based prejudices). In the best case scenario (assuming there is no duplication of questionnaires) and their number is close to that of the recipients

of social services) such survey could be representative just for the recipients, not for the whole ethnic group. This is something at least.

Community-based monitoring

Community-level data is particularly important in regards Decade progress assessment – and this is the information that is usually underrepresented in standard statistical instruments. This is why the possible approaches for generating ethnic-sensitive data outlined above should be complemented by comprehensive system of quantitative and qualitative data collection at community level. Such a system should provide basic information on the communities in question based on standard questionnaires filled in on regular basis by designated member of the community after receiving training on basic data collection and reporting techniques. The system would provide:

- Quantitative information on the community status (number of households, their housing conditions, number of children attending school, their age and grade, number of drop-outs, number of new-born, number of vaccinated children etc.).

- Quantitative information on occurrence of certain events relevant from Decade monitoring perspective (power cuts and their duration, accidents, conflicts with majority or other Roma groups, NGOs activities etc.).

The data collected within the system of community monitoring will provide data on the status of Roma communities, their internal dynamics and the life in Roma neighborhoods, particularly in the closed ghettoes. In this regard they will provide data that will be complementary to other sources. For complementarity purposes its structure of data (and the design of the instruments used) should be as close as possible to other instruments for similar data collection. Necessary precondition of training of the local data collectors on basic data collection techniques and standards and establishing a system of incentives for responsible and reliable work as well as control system. Given the necessary investments, establishing and putting into operation of such a system would require time (a year at least).

There are major problems in this regard however. The Roma communities are "interested party" and data generated within such methodologies can be complementary to the data from other sources. Local monitors can be under pressure from local leaders, they may be lacking the necessary qualifications and just a single training may be insufficient to offset these deficits. But the lack of "common interest" spirit ("us versus them" phenomenon) may emerge as a major problem.

Despite these potential (and probable) problems, efforts in community-based monitoring should be supported. Working in this direction is in fact a major contribution to overcoming the problems outlined above.

Putting data in context – secondary source contextualization

Achieving tangible progress in the priority areas of Roma inclusion is a major objective of Roma inclusion efforts. Local specifics—available employment opportunities, environmental degradation, local conflicts, extent of participation—often determine opportunities in this regard. These parameters of the local context can be quantified and reflected in the process of (and policies targeting) Roma inclusion.

The underlying assumption of this approach is that individual characteristics (grasped in indicators based on survey data) have various determinants. Correlating the individual characteristics with those determinants may reveal important links. They do not necessarily entail causality but applying in-depth research and qualitative methods those causalities can be determined as well.