

Work Session on Statistical Metadata
(Geneva, Switzerland, 18-20 February 1998)

Item 3 of the provisional agenda

**REPORT ON PROGRESS ON THE HARMONIZATION
OF SOCIAL STATISTICS**

Submitted by Statistics Canada ¹

¹ Prepared by G. Priest.

I. INTRODUCTION

1. Recent initiatives at Statistics Canada related to meta information have been reported at a number of meetings over the past two years.² The first in the series, presented at the American Statistical Association, set out the origins and nature of problems related to the lack of meta information and the lack of harmonization. The second paper, presented at the CES workshop in Berlin, reported on work that was underway in addressing the matter of meta information. The third paper, presented at the Eurostat Workshop on Harmonization in London, proposed how the issue of disharmonies between statistical sources might be addressed.

2. To summarize those papers, it may be said that it has been recognized that information technology has bred a new generation of increasingly sophisticated and demanding users of statistical information. However, that same technology has also facilitated the building and maintenance of comprehensive meta information systems as it has provided the tools to begin effectively addressing the issue of conceptual disharmonies between sources. Indeed, the first step in addressing disharmonies is to consult with clients in order to assess the implications of those disharmonies in terms of how effectively data may be integrated and utilized. It is also a matter of making some assessment of the cost implications of unnecessary and redundant work in program areas (because of the failure to use shared concepts and classifications). The second step is building a corporate meta information system in which resides documentation related to concepts, definitions, classification systems and collection and processing systems. It is this meta information which must be studied in order to determine the existence, the extent and the nature of any disharmonies.

3. At Statistics Canada a browsable corporate meta information system has been built for social statistics and it is under development for other fields. A template has also been developed which will electronically and seamlessly harvest such information from new or ongoing statistical activities as they are developed.³ The meta information in this system has proven invaluable to staff and clients searching for sources of information. Of equal importance, however, it has been indispensable in facilitating work on harmonization.

4. Conceptually disharmonious data are generally manifested in one of two ways. First, there may be cases where data actually are comparable from source to source but different naming conventions are used. For example, we determined that some sources had a variable named, *educational attainment*, while other sources had a variable named, *highest level of schooling*. The two were conceptually the same but the different naming conventions

² See the following:

“In Search of Data Integration: No Matches Found,” Gordon E. Priest, **1996 Proceedings of the Section on Survey Research Methods**, American Statistical Association, Chicago, August 1996.

“Report on Progress in Implementation of Statistical Metainformation Systems,” B. Slater and G. Priest, Work Session on Statistical Metadata, Conference of European Statisticians, Berlin, October 1996.

“The Issue of Harmonization of Data from Diverse Sources,” Gordon E. Priest, Eurostat Workshop, London, November 1996.

³ See the companion paper from Statistics Canada by E. Boyko.

confused clients and diminished the capacity to effectively integrate data from the different sources. Second, and a much more serious implication of disharmony, are cases where data use the same naming convention from source to source but are, in fact, conceptually different. In such cases, clients may use the data improperly, unaware of the conceptual differences. An example of this type of disharmony was manifested in the concept of *marital status*. Some sources pertained to *legal marital status*, others pertained to *de facto marital status* and yet others combined *legal* and *common-law status* but they all used the simple term *marital status* to name the variable.

5. Even when data are conceptually harmonious from source to source, differences in classification strategies may make data, largely or in part, difficult to integrate. An example in this case relates to a variable titled *census family structure* where some sources aggregated the data into subcategories related to the marital status of the spouses. Other sources aggregated into subcategories related to whether children in the family were natural children or step children.

6. The identification of these kinds of disharmonies requires a painstaking review of the meta information as well as output programs. The latter will generally quickly reveal differences in classifications used in publications and even public use micro data files. They may or may not reveal conceptual disharmonies, depending upon how well the output is documented. More revealing is a review of the classification systems used in what we call the retrieval micro data files from which all outputs, including both standard products and custom tabulations, are produced. Obviously, documentation related to concepts and definitions must be reviewed as well. In addition, sometimes a level of ambiguity remains such that a detailed review of collection and processing specifications is required to determine the extent and nature of potential disharmonies. That is, there may be cases where there is apparent harmony but differences in editing, imputing or weighting may introduce disharmonies.

7. The objective of the review is, of course, to not only identify the disharmonies but to resolve them. Where variables from different sources are conceptually the same then they should use the same naming convention. Where they are different, different naming conventions should be used. For any given variable, a range of standard classifications should be used. They may be highly detailed or they may be highly collapsed but they should always *map* to one another in terms of subcategories that may be used. For example, if a collapsed classification of the variable *class of worker* shows only *paid employees* and *self-employed*, then these two categories should also be reflected in a highly detailed classification which shows additional dimensions such as *employed in the private sector* and *employed in the public sector* or the self-employed *with paid help* and *without paid help*.

8. The resolution of the disharmonies is often facilitated by identifying the practice of one of the sources as the *best practice*. Sometimes, however, none of the sources is completely satisfactory and a new best practice must be identified. Where a variable is derived from diverse sources, including administrative records, additional constraints may be in place. Where data are derived from the census or surveys within a single agency, it is in the power of the agency to force harmonization in terms of how the data are collected and processed. However, in the case of the derivation of data from administrative sources the degrees of freedom to influence data collection, and perhaps data processing, are greatly diminished. Therefore, if it is desirable to ensure harmonization between administrative sources and survey sources, it may be necessary to adjust the collection and processing of the latter. That is, there may be cases where a variable is conceptually driven by the limitations of administrative sources. It should be added at this point, that there will be cases where differences in survey coverage or methodology from source to source may precipitate constraints or force compromises, even for single agency sources.

9. It must be pointed out that the intention in harmonization is not to force fit data collections to a *one size fits all* concept and classification system. It is recognized that clients may have legitimate reasons for requesting different measures or different perspectives on a given characteristic. A statistical agency must respond to those discrete client requirements but it needs to ensure that naming conventions are adopted that clearly reflect the different perspectives.

10. Once best practices have been identified it then remains to undertake a broad consultation. Certainly, the program areas collecting any given variable need to be consulted to determine whether they can indeed implement the proposed practice. Clients need to be consulted to determine whether the proposed practice meets their needs. It is also desirable that international consultation takes place to better facilitate international harmonization. This implies the need for an iterative process whereby the diverse partners in the process can be cognizant of the many revisions which are likely to take place as the process of consultation and negotiation moves forward.

11. The challenge at this point becomes one of document control and management. The best practices which are being recommended as potential standards need to be put into a document that meets the following criteria:

- browsing and navigation must be facilitated by both a thematic index and a keyword search engine;
- since many revisions to the document are possible it is necessary to isolate and date those parts of the document where revisions are made;
- the partners in the process need to know who are the other partners and who is being consulted;
- the partners need to know the range of sources for any given variable, that is, the statistical activities which are currently collecting data for the variable;
- the partners need to know the status of the work, that is, is the current state of work such that it remains a recommended best practice under discussion, is the state such that discussion is finished and the work is being recommended to the corporation for acceptance as a standard or has it been accepted as a standard?

12. With respect to the content of the document, there is a need to present the following:

- discussion of broad concepts, for example, *marital status* may be considered as a concept to which a number of variables such as *legal marital status*, *common-law status* and *conjugal status* are related;
- recommended definitions which textually describe a variable by clearly stating its meaning and what it includes and what it does not include;
- recommended classification systems which reflect both the most detailed array for a variable as it would exist on a micro retrieval file as well as truncated or collapsed arrays which might appear in an output.

13. It was immediately concluded by Statistics Canada's social statistics harmonization team that paper documentation was unsuitable and that a networked electronic document was necessary. The document needed to be accessible through the agency's Internal Communications Network (ICN). It was also deemed desirable to have it accessible to external partners via the Internet. As a result, **Folio** was selected as the software of choice for a number of reasons. It had been used on the ICN for a number of years, the above-noted meta

information document had been built using it and methods existed to convert **Folio** documents to *html* on the fly.⁴

14. Work began on the harmonization initiative in April 1997. It built on some work done the previous year on *age*, *marital status* and *absence from work* variables as well as some work done on *family* variables two or three years earlier. While the state of existing documentation and the availability of experienced resources was a factor in determining the areas in which to work, priority was given to those variables which are sometimes identified as *classification variables* because they are used for the socio-economic classification of the population in most censuses and household surveys. There were two reasons for this choice. The first was the influence of the Eurostat workshop on harmonization held in London in November 1996. The other was that Statistics Canada was about to commence a major client consultation related to the 2001 Census which would provide a forum for testing recommendations with major client groups.

15. The first task undertaken was the organization of the document, or rather documents, since it was decided to have separate English and French versions with hypertext links between them. It was also decided that hypertext links to the meta information documents were necessary to facilitate identifying sources of the variables and the meta information connected to those sources. Content was divided into themes such as *demography*, *education*, *environment*, *ethno-cultural* and so on. Each theme was divided into subjects. For example, in the case of *demography* identified subjects included *age*, *fertility*, *gender*, *marital status* etc. Each subject was to have a textual concept attached to it which explained relationships between the subject's associated variables. Each subject also included a list of its associated variables. For example, *marital status* included the variables *legal marital status*, *common-law status*, *conjugal status*, *census family status*, *economic family status*, *household maintainer status* and *household status*. This organization is reflected in an electronic index which can be toggled to search at the theme level, the subject level or the variable level.

16. The second task undertaken was to develop a model and standards for the presentation of each variable. It had been noted that no standards existed for the presentation of definitions. The definition of variables related to the population sometimes made reference to an individual or to individuals, sometimes to a person or persons, sometimes to people and sometimes to the population. It was decided that *all* definitions related to the population should refer to a *person*. For example, "*Conjugal status* refers to whether a person aged 15 and over is living with a person of the opposite sex as husband, wife or common-law partner." The use of the term *population* was retained for the classification systems where reference is made to the *total population* displaying a particular characteristic. Thought was given to using the term *total persons* but it was felt that the former was more commonly used and understood.

17. Finally, work began on review of the meta information related to the variables and the conceptual development of recommended practices. This work was done simultaneously in both the English and French documents. The prime reason for this was that previously, concepts, definitions and classification systems had often been done in one language and then *translated* into the other which in itself precipitated some linguistic disharmonies. By the first of July, a sufficient amount of work had been done to post the documents to the ICN in order

⁴ Some caution is in order in this regard since the conversion process modifies some of the format and font settings and some experimentation is required to resolve the problem. It is expected that this minor problem will be eliminated in forthcoming releases of the software.

to begin consultation with the program areas and relevant committees. By the end of the month the documents had been posted to a password-protected Internet site to facilitate consultation with external clients and partners. The site was password-protected to avoid potential confusion amongst clients with regard to what was being recommended and what was actually currently in use. The URL and password can be obtained on request.

18. While paper copy does not do justice to the ease of working with the electronic document, an example of how a *subject* appears in the document is shown in *Figure 1* (see Annex).

19. The example of the subject of *marital status* shown in *Figure 1*. states the term in English and then in French. The French term is highlighted as a hypertext link and clicking on it moves the viewer to the same place in the French document. Opening the latter as a second window, both English and French documents can be viewed at the same time thereby facilitating a comparison of the English and French concepts. In this example, the concept is quite straightforward but in others, such as *relationship between household members*, it is quite detailed and extensive. The *consultation status* shows how far the task of consultation has progressed. In this case, the program areas have been consulted and their requested revisions negotiated and incorporated. Relevant subject matter committees,⁵ however, are still reviewing the work but the relevant advisory committees have not yet discussed it. Consultation is in progress with relevant federal and provincial committees and with international bodies but consultation with major clients is still pending.

20. *Figure 2* (see Annex) shows the documentation for a typical variable. In this case the example is the variable *census family status*.

21. The example, *Census family status*, is a variable related to the subject of *marital status*. The French term for this variable is *Situation dans la famille de recensement*. This is also hypertext-linked to the French document. Note that the definition makes reference to a *person* in the singular. In this case, not only is the variable defined, so are the response categories since there has been some ambiguity in the use of these class intervals in the past. The detailed classification reflects the most detailed classification collected for this variable but it also reflects the two subcategories, *census family person* and *non-census family person*, which are used in the collapsed classification. This permits comparison between sources that produce detailed outputs and those which are only able to produce collapsed outputs. An optional classification might have also been expressed. For example:

Total population
 Census family person
 Husbands, wives and lone parents
 Children and youth
 Non-census family person

⁵ Within Statistics Canada, some committees have been struck to help improve conceptual harmonization and integration of data between statistical programs. For example, there is one for family and household statistics. There are also external advisory committees. For example, the Advisory Committee on Social Conditions provides advice to the agency on statistical programs which produce data related to social conditions. Advisory committees primarily represent clients from the academic and business sectors and provide advice with respect to adjustments to current programs or to the development of strategic plans. Committees comprised of federal government departments and provincial statistical agencies provide similar advice for the public sector.

The key point here is that the subcategories, *census family person* and *non-census family person* are retained to ensure cross-source comparability.

22. Other information provided includes the reference to *International concordance*. Here it is hoped to document those cases where the variable is in concordance with standards used by other national or international agencies. *Source(s) for variable* provides a hypertext link to the meta information base to determine which statistical programs collect data for the variable in question (see Figure 3 in Annex). *Working status* indicates whether this variable is still a *recommended best practice* under discussion or whether the discussion and negotiation is complete and it is a *recommended standard* awaiting final approval by the departmental Methods and Standards Committee. *Standard* indicates that the definition and classification have been approved as a departmental standard. *Date of last revision* is an important reference which reflects whether revisions have been made, and if so, when. *Earlier version* provides a link to an archive of earlier versions of the variable that have since been revised or amended.

23. *Figure 3.* reflects the listing of sources of the variable, *Census family status* in the meta information base.

This indicates that three statistical programs have produced the variable, or some variation of it for the dates shown. It does not indicate whether or not the output in the past has been harmonized. It may have been in full, in part or not at all. Nevertheless, this link is critical because it identifies sources that had to be checked. It also provides access to key contacts in the program areas and to such meta information as exists in electronic form. The dates are all hypertext linked to meta information for each particular source as shown in *Figure 4* (see Annex).

24. *Figure 4.* displays the range of meta information available for the 1991 Census of Canada. Each of the titles is a hypertext link to various fields of meta information. *General information about this survey* contains basic information about methodology such as the type of statistical activity, the frequency of collection, the sample size, geographic coverage and detail and output media. *Concepts and terminology* provides a link to electronic dictionaries (where they exist). *Overview of data quality* provides notes on measures of data quality for the source (where it exists). *Questions asked and overview of data collection* provides a link to electronic questionnaires where available. *Themes and subjects covered* is a link to a thematic list of all the variables (direct and derived) available from the source. *Public Use Micro Data Files* provides a link to record layouts or data dictionaries available for those sources that produce public use files. *Contents and layout of main data base* provides a link to record layouts for the master retrieval data bases from which all outputs are produced. *Products and Services* provides a link to a catalogue which displays references to electronic and paper products produced from the source. *For more information* is a link to a program manager who can provide more information about the particular source or provide access to documentation which may not exist in electronic form suitable for distribution.

25. These links to the meta information base were invaluable in the review stage of the harmonization work in terms of determining which sources had, in the past collected data on a given variable. Where electronic meta information had been ported to the meta information base it could be quickly located and browsed through keyword searches: something which is much more difficult to do with uncatalogued paper documentation. Nevertheless, where electronic documentation was not available we at least had the names of contact persons from whom we could obtain paper documentation or seek clarification. Once a review was done and recommendations made, this was also used to identify the program areas and the managers with whom the recommendations had to be discussed and negotiated..

26. At this point, recommended definitions and classification systems have been produced for about 200 variables of which some 30 or 40 might be considered important classification variables. We are concentrating our efforts on the latter in reaching agreement between the program areas and, of course, it is the latter on which it would most desirable to reach international agreements. While we are anxious to bring closure to the work we do realize the necessity of consultation, not only with the program areas but with client groups as well. In this regard our experience has been that there are pros and cons to such consultation. On the one hand, it is critical to ensure that client needs are met and that what we propose can be implemented. On the other hand, consultation poses some risks. It can raise unreasonable expectations beyond what can reasonably be done. Consensus is very difficult to obtain and it is very difficult to bring closure to the discussion.

27. Harmonization also bears a price, both in terms of resources and in terms of potential disruptions to long-standing historical series. The review of the meta information and the development of recommendations has necessitated the use of knowledgeable experienced staff who are generally at the upper end of the pay scale. Therefore, there has been a relatively heavy monetary expenditure in the work. Harmonization will force a break in some historical series and some further resource expenditure may be necessitated in some programs in terms of the development of explanatory notes which will assist clients in mapping old outputs to new outputs.

28. Cost considerations have played a role, along with other reasons, for not, at this time, attempting to develop harmonized or standardized questions. It was reasoned that harmonization of definitions (the statement of what a variable includes and what it doesn't) and classifications was more important than the harmonization of questions. Clients do not use questions, they use outputs. The harmonization of questions does not deal with the issue of the frequent need to harmonize with administrative sources. And the harmonization of questions may not deal effectively with the production of derived variables. Furthermore, experience has shown that harmonized questions do not necessarily yield harmonized results. Differences in interviewing methodologies as well as differences in environment or context may yield different results to the same question. Therefore we have argued that program areas must adopt the standard definitions and classifications systems but they must be given some degree of freedom to pose their questions according to the constraints of their methodology and collection vehicle...as long as their outputs map to the standards.

29. Of course, it remains to be seen whether this works in practice. There also remains a concern that program areas may not take, or not be able to take, an harmonious approach to the processing of data. Indeed, consistency in editing, imputing and weighting may be compromised due to differing collection methodologies. For example, editing and imputing of missing or inconsistent values can be enhanced where computer assisting interviewing is in use since a considerable amount of verification can be done during collection. Even with the best of intentions, it may not be possible to replicate that kind of thoroughness when paper and pencil collections methods are used. Furthermore, editing and imputing strategies are very much influenced by content. That is, the range of related variables included that can be used for verification will have an impact on edit and imputation strategies from program to program.

30. Thus, the emphasis of the work to date has been to drive for the harmonization of definitions and classification systems. Nevertheless, the door is not closed to the possible need to do further work which would develop guidelines, if not standards, for collection and processing systems, including modules of questions and editing and imputing strategies. There may also be benefits in standardizing coding structures for data dictionaries and record layouts.

31. With regard to future directions for the initiative there remains the issue of whether the work in the social field should be integrated with corporate initiatives. A long standing corporate policy on the documentation of data quality and methodology does stipulate that data releases must be accompanied by an explanation of fundamental concepts, including basic definitions. It also states, “...if similar data from other important sources exist, these sources should be identified. Where appropriate, a reconciliation should be attempted and a description of how the data sets differ and the reasons for these differences given. Comments on the quality for the other data set should be provided, if an evaluation is available and relevant.”

32. In general, there has been compliance with the requirement to include definitions although the clarity and quality of the definitions themselves is very uneven from program to program. Statements of sources of other related data sets and statements on comparability, on the other hand, have been almost non-existent...in spite of the policy. While the agency has a Methods and Standards Committee and a Standards Division, as well as a host of subject matter committees, there has been little or no monitoring with a resultant high level of non-compliance. Perhaps, in fairness, it should be said that until the meta information base was developed, program areas often knew very little about the collections of other program areas. How could they make statements of comparability concerning sources of whose existence they were not aware? On the other hand, had the extent of some of the disharmonies been revealed perhaps the hue and cry from clients might have arisen to an even greater degree than it has?

33. At any rate, it is hoped that through the current initiative such disharmonies will be resolved. But it remains to be determined whether this process, developed within social statistics should be a process developed at the corporate level. Should agreements struck within social statistics be endorsed by the Methods and Standards Committee? We have assumed that they would be but there remains the issue of how standards might be promulgated in an authoritative manner such that program areas clearly understand that compliance is compulsory. And there is the issue of how the program areas are to be monitored to ensure compliance.

34. These are issues for future consideration that may move the initiative from the social statistics field to the corporate level.

35. In summary, the cornerstone of the work to date has been the prerequisite development of a meta information base which provided access to the documentation that had to be reviewed to not only identify disharmonies but permit the development of recommended best practices. Extensive consultation is necessary and the work has focused on the development of harmonized definitions and classification systems. There may, or may not, be need to do further work on the harmonization of collection and processing specifications. There may, or may not, be a need to move the initiative to a corporate level.

Figure 1.**Subject: Marital status***État matrimonial***Concept:**

The concept of *marital status* applies to the conjugal or living arrangements of persons. Seven types of *marital status* are identified: *legal marital status, common-law status, conjugal status, census family status, economic family status, household maintainer status* and *household status*.

Consultation status:

Statistical activity program areas: completed

Subject matter committee: in progress

Advisory committee: pending

Relevant federal/provincial committee: in progress

Major clients: pending

International bodies: in progress

Figure 2.**4. Census Family Status***Situation dans la famille de recensement***Definition:**

Census family status refers to a person who is co-resident in a household and whether he or she is a member of a census family, and if so, his or her role in the family. A *husband* or *wife* is defined on the basis of living with a *spouse* while a *lone parent* is defined as living with his or her own *child* or own *youth*. In this context a *child* is defined as a person aged 0 to 14 co-residing with one or more parents or guardian. A *youth* is defined as a person aged 15 to 24 who is co-residing with one or more *parents* but not with his or her own *spouse* or *child*. A *Youth* who is co-residing with one or more parents but who is also co-residing with his or her own *spouse* or *child* is considered as a *husband*, *wife* or *lone-parent* in his or her own right.

By definition, all persons who are members of census families are also members of economic families. The total number of persons in economic families, therefore is the count of persons in census families plus the count of non-census family persons who are members of economic families.

Detailed classification:

Total population

Census family person

Husband or wife (legally married spouse or common-law partner)

Lone parent

Child

Youth

Non-census family person

Member of economic family

Not member of economic family

(Continued next page)

Figure 2. Continued**Collapsed classification:**

Total Population

Census family person

Husband or wife (legally married spouse or common-law partner)

Lone parent

Child

Youth

Non-census family person

International concordance with:

Under investigation

Source(s) for variable:

[link to meta information base]

Working status:

Recommended best practice x

Recommended standard

Standard

Date of last revision: 09:09:97

Earlier version

Figure 3.**Meta information base (thematic search tool)****Census family status***Census of Canada, 1986, 1991**Survey of Consumer Finances, 1992, 1994, 1995**Household Income, Facilities and Equipment Survey, 1991, 1992*

Figure 4.

1991 Census of Canada

General information about this survey

Concepts and terminology

Overview of data quality

Questions asked and overview of data collection

Themes and subjects covered

Public Use Micro Data Files (Technical characteristics and record layouts)

Contents and layout of main data base

Products and Services

For more information