WP.6
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Workshop on the Common Metadata Framework
(Vienna, Austria, 4-6 July 2007)

Topic 2: Functions and attributes of statistical metadata

THE IMPORTANCE OF METADATA IN THE USE OF ADMINISTRATIVE SOURCES

Submitted by the Katholieke Universiteit Leuven

An Taelemans (Centre for Sociological Research)
Mieke Booghmans (Policy Research Centre Work and Social Economy)
Jos Berghman (Centre for Sociological Research)
Luc Sels (Policy Research Centre Work and Social Economy)

INTRODUCTION

1. 'Metadata are as important as data and need as much work as data' (Lindblom & Sundgren, 2004, p. 17). This rule, suggested by Statistics Sweden, underlines the importance of treating data and metadata on an equal basis, especially in the use of administrative data.

2. The Belgian authorities support the increasing use of administrative data but also acknowledge metadata as indispensable in this respect. That is why Belgian Federal Science Policy Office is currently financing a research project with the explicit aim to create high quality and comprehensive metadata that accompany the administrative data stored in the Datawarehouse Labour Market and Social Protection (Datawarehouse LM & SP).

3. This paper describes the developing process of a metadata strategy for the Datawarehouse LM & SP. In the first section an overview is given concerning the history and construction of the Datawarehouse. Then the current research project and its objectives are situated. While developing a metadata strategy, three important questions have arisen. Firstly, ***why*** are metadata important, secondly ***what*** do we expect of these metadata and finally ***how*** should the metadata be collected, stored and disseminated in a structured and efficient way? These questions are answered in the third section of this paper. The following section focuses on the benefits of metadata by elaborating examples that illustrate the necessity of comprehensive metadata, especially in the use of administrative data. The final section explains how the presented metadata strategy is implemented and put into practice.

A. DATAWAREHOUSE LABOUR MARKET AND SOCIAL PROTECTION

4. The Datawarehouse Labour Market was set up in 1998 at the request of several social security institutions and scientists.[1] The objective was to create a database in which administrative data of social security institutions would be permanently stored in order to be more readily accessible for scientific research. In the following table the social security institutions that participate in the Datawarehouse Labour Market and Social Protection are listed as well as the years for which their administrative data are integrated.

Table 1. Social security institutions that participate in the Datawarehouse Labour Market and Social Protection

	1997	1998	1999	2000	2001	2002	2003	2004	2005
National Service for Medical and Disablement Insurance	X	X	X	X	X	X	X	X	X
Department of Social Services responsible for child benefit	X	X	X	X	X	X	X	X	X
National Institute for the social security of the self-employed	X	X	X	X	X	X	X	X	X
National office for social security	X	X	X	X	X	X	X	X	X
National office for social security for local and provincial authorities	X	X	X	X	X	X	X	X	
National employment office		X	X	X	X	X	X	X	X
National Health Services							X		
Industrial Accidents Fund							X		
Occupational Diseases Fund							X	X	X
National Office for Social Integration, Poverty Prevention and Social Protection							X	X	
National Office for Pensions							X	X	X
National Register	X	X	X	X	X	X	X	X	X

[1] For more information about the history and the applications of the Datawarehouse, we refer to: C. Vermandere, A. Vanheerswynghels & P. Van der Hallen (2006). *Eén plus één is drie. Datawarehouse Arbeidsmarkt en Sociale Bescherming: verleden, heden en toekomst (One and one equals five. Datawarehouse Labour Market and Social Protection: past, present and future).* Gent: Academia Press.

5. Table 1 shows that since 2003 new administrative sources related to social protection, were integrated into the Datawarehouse. Therefore the scope of the Datawarehouse Labour Market expanded and resulted in a change of name, namely Datawarehouse Labour Market **and Social Protection**.

6. The population of the Datawarehouse firstly contains all persons known by at least one participating social security institution. The household members of these persons are added to the population via the National Register, insofar these household members are not already known by one of the participating institutions.[2] It is clear that the more social security institutions participate in the Datawarehouse, the higher the coverage rate of the Datawarehouse will be in terms of the total Belgian population. At this moment the Datawarehouse covers approximately 98 percent of the total Belgian population.

7. The administrative datasets are stored in the Datawarehouse by quarter. Several social security institutions already supply their datasets quarterly. In case the institutions supply annual or monthly datasets, the administrative data are assigned to a quarter by means of the starting and final date of the reference period.

8. The Datawarehouse LM & SP has three advantages. Firstly, it provides *ready access at a minimal cost* to data on the socio-economic position of the population. Secondly, it becomes feasible to generate wide-ranging and detailed *statistics* about socio-economic operations. In order to make these data easily accessible, a few basic statistics were compiled with the most frequently requested variables to illustrate the Datawarehouse's possibilities. These basic statistics are available to the public. Specific (tailor-made) statistics are obtainable on request. Thirdly, the Datawarehouse makes it possible to link the administrative data of the participating social security institutions. The basis for this link is the national insurance number, a unique identification number held by everyone known to the Belgian social security institutions.[3] An individual is therefore the main *statistical unit*.

9. Each participating social security institute offers an extensive set of *variables* through the Datawarehouse. In addition, a number of *derived variables* is created. The socio-economic position is a commonly used derived variable which makes it possible to divide the population according to their position in or out the labour market.[4]

---

[2] If the household members are not known by the participating social security institutions, the National Register offers only information on the gender, address, head of the household and the relation to the head of the household. Nevertheless, this information is sufficient to create a typology of households (LIPRO). LIPRO (Lifestyle Projections) is a typology of housesholds that is frequently used for demographic research in Europe.

[3] The national insurance number is coded and made anonymous in the Datawarehouse

[4] The situation taken into account is invariably the situation on the last day of the quarter.[4]

10. The nomenclature of these socio-economic positions is constructed hierarchically and it can be distributed to a three-digit level.

Table 2. Nomenclature of socio-economic positions

1. Employed

1.1. Wage-earner
1.1.1. Wage-earner with one job
1.1.2. Wage-earner with more than one job
1.2. Self-employed
1.2.1. In main occupation
1.2.2. In side occupation
1.2.3. Self-employed above retirement age
1.3. Working as assistant of self-employed employer
1.3.1. Working, main occupation as assistant
1.3.2. Working, side occupation as assistant
1.3.3. Working as assistant after retirement age
1.4. Wage-earner as well as self-employed
1.4.1. Wage-earner in main occupation
1.4.2. Self-employed in main occupation
1.4.3. Assistant in main occupation

2. Jobseeker
2.1. Jobseeker following full-time employment, receiving benefits
2.2. Jobseeker following voluntary part-time job, receiving benefits
2.3. Jobseeker following studies, entitled to waiting allowance or bridging grant
2.4. Jobseeker with guidance allowance

3. Professionally inactive
3.1. Full-time career break
3.2. Exempted from reporting as jobseeker
3.3. Welfare benefit/Financial assistance
3.3.1. Welfare benefit
3.3.2. Financial assistance
3.4. Pensioner without work
3.5. On full-time bridging pension
3.6. Children giving the right to child allowance
3.7. Disabled

4. Other

B. SETTING OF THE RESEARCH PROJECT

11. As from October 2005, both the research team of the Centre for Sociological Research and the Policy Research Centre Work and Social Economy are engaged in a four year project, called 'Datawarehouse Labour Market and Social Protection: Expansion concerning content and methodology'.[5] This research project is financed by Belgian Science Policy and is under the authority of the Federal Public Service Social Security and Crossroads Bank for Social Security.

12. To achieve a comprehensive Datawarehouse Labour Market and Social Protection, it is necessary to widen its scope by opening up new datasources or by expanding the existing ones. To keep the Datawarehouse up to date, also more recent data of the social security administrations have to be included on a regular basis.

13. While widening the scope of the Datawarehouse, priority has to be given to a systematic quality check of the administrative data, not only at the technical level but also with regard to the contents of the available data. The quality has to be guaranteed so that the administrative data are correctly used and interpreted. The main objective of the research project is to achieve a consistent, harmonised and high quality Datawarehouse Labour Market and Social Protection.

14. Because the current Datawarehouse Labour Market and Social Protection faces problems of fragmented, inconsistent and insufficient metadata on the one hand and problems of harmonisation on the other hand, our research project has the intention to create comprehensive and high quality metadata and to make recommendations for the further (inter)national harmonization and optimization of the Datawarehouse. To attain this objective, three important movements can be distinguished: a descriptive task, an analysing task and a task of evaluation. These three movements interact and cannot be seen as separated tasks.

15. The ***descriptive task*** implies that an adequate description has to be made concerning the composition and contents of the datasets of the social security institutions, as well as a clear documentation of the way these datasets are integrated into the Datawarehouse. Once this rather descriptive information is available, the focus has to shift to an ***analysing task***. Because it is possible, even probable, that different administrations give similar interpretations to differently operationalised concepts and variables, a systematic and thorough screening of the concepts and variables used in the datasets has to be made. In that way documentation and metadata can be an instrument of harmonization and comparability. Also international harmonization is an issue that is recognized in this research project through the third ***task of evaluation***. The research team has

[5] The Policy Research Centre Work and Social Economy is accompanied by a third partner, namely the Point d'Appui Travail Emploi Formation.

the task to evaluate the administrative data in a European/international perspective and make recommendations to harmonize the Belgian data with international standards. The Social Security Matrix is a useful instrument for this task of evaluation. The matrix indeed indicates which information is needed for an optimal Datawarehouse Labour Market and Social Protection and takes into account the international standards and definitions. It is a framework that constitutes the basis for an extensive, longitudinal database by which a new and better picture on Belgian social protection can be obtained (Nijs & Berghman, 2005, pp. 1-2).

C. DEVELOPMENT OF METADATA STRATEGY

16. At the beginning of this research project, three main questions arose. First of all: ***why*** are metadata important, secondly ***what*** do we expect of these metadata and finally ***how*** should the metadata be collected, stored and disseminated in a structured and efficient way? These questions are an excellent basis for putting priorities and establishing a metadata strategy.

a. Why are metadata important?

17. Administrative data are a major source of information for government, public and researchers and can provide substantial advantages in terms of cost and respondent burden. Because administrative data were not originally compiled or produced for statistical purposes, it is particularly important for such data to be methodologically transparent. An assessment of the administrative records should be made in terms of coverage, content, concepts and definitions, timeliness, frequency and quality.

18. Metadata are not only a part of but also a condition for quality. Documentation contributes to the interpretability of the administrative data and provides the means of assessing data quality. Quality can be defined along a number of dimensions, but there is no international standard nor universally accepted quality model. There is however rather good convergence among the different quality frameworks. All the aspects that affect the data quality should be covered by comprehensive metadata.

19. Certainly when it comes to working with data from different administrative sources, metadata can also be used as a tool for harmonizing concepts and definitions. Because administrative institutions can define variables and concepts in different ways, a problem of harmonisation arises. Metadata can identify these problems so that recommendations can be made towards harmonisation and comparability.

b. What do we expect of metadata?

20. Before creating metadata, it is important to specify the criteria that our metadata should meet. We determine four key functions. Metadata should be:

- *Flexible*

The Datawarehouse LM & SP entails a broad variety of users, such as researchers, civil servants of the social security administrations, local governments, etc. Therefore our metadata should meet the needs of all these different users, ranging from the 'general' to the 'professional' user. A flexible and user oriented approach also implies that contact details of experts are available and can be used in case further information is needed.

- *Comprehensive and consistent*

Metadata should provide a complete, precise and unambiguous description of the administrative databases and variables. This documentation should cover various dimensions: technical information on the observation unit and structure of a dataset, information concerning the content of the variables and information on the purpose for which the administrative data were originally collected and the different ways in which the data might be transformed by the providers.

- *Up-to-date*

Stability is an important concern. Therefore changes in the regulation and in the method of data recording and processing by the social security institutions have to be monitored. The impact of these changes on the administrative data should be included in the documentation. Sometimes however it is not sufficient to keep the metadata up-to-date. The historical versions of the metadata have to be maintained as well in order to correctly interpret the administrative data of the past (Sundgren, 2003, p.24).

- *Accessible*

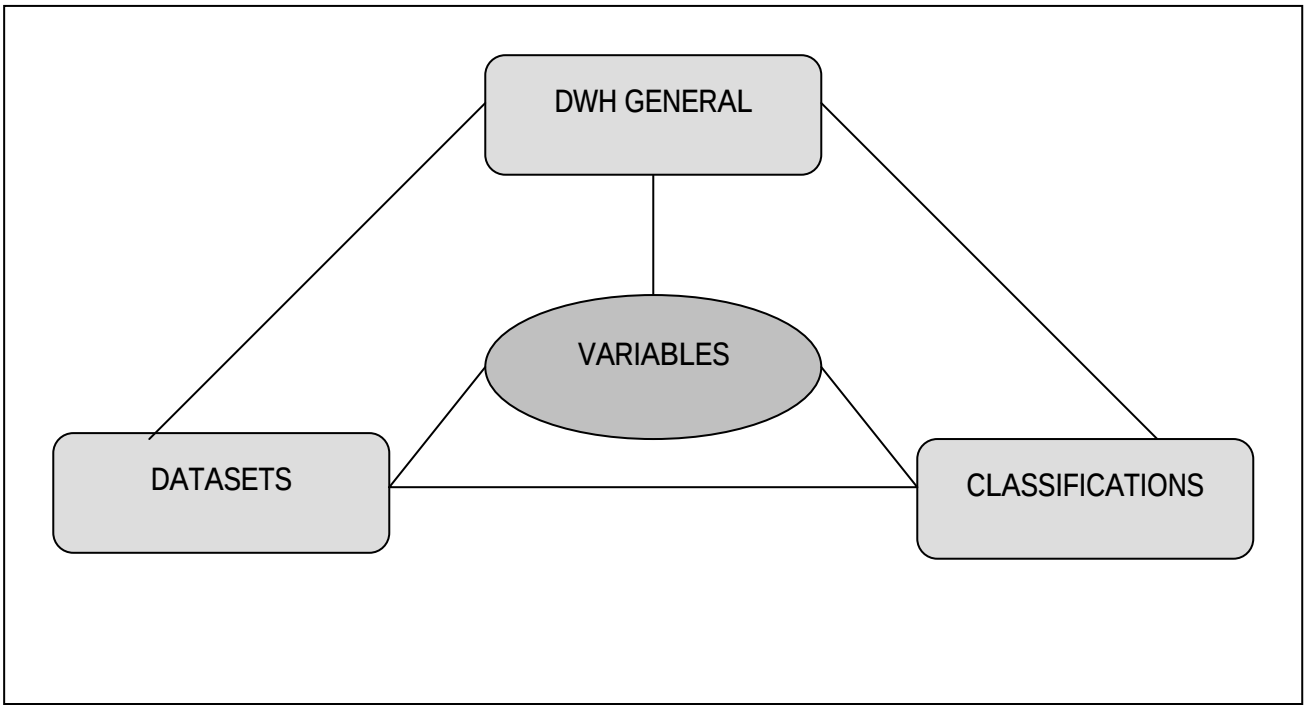
Ideally, metadata should be directly linked to the data they describe. Because of privacy concerns however, it is not possible to make the administrative data of the Datawarehouse publicly accessible. An application is required to receive the data. This does not mean that metadata that are not directly linked to the administrative data they describe, are useless. On the contrary, by making metadata publicly available, (potential) users receive information on which data are available and on the possibilities and restrictions of these data. Internet is the best medium for obtaining up-to-date metadata.

c. How should the metadata be structured?

21. It is extremely important that the users of the Datawarehouse LM & SP have easy access to the metadata accompanying the administrative data. The large amount of available metadata increases the need for efficient navigation and search. Therefore it is a major challenge to collect, store and disseminate this high amount of information in a well structured and harmonized way.

22. To develop such an IT structure, we started by making a benchmarking exercise. An inventory of existing metadata systems was made in order to find a system that could meet the postulated criteria. On this basis a variable-centred metadata model for the Datawarehouse Labour Market and Social Protection was created.

Figure 1. Metadata structure of the Datawarehouse Labour Market and Social Protection



23. An approach commonly used, entails the presentation of metadata as layers within a pyramid, progressing from summary metadata on top of the pyramid to more detailed metadata at the bottom (OESO, 2006, pp. 73-75). This layered presentation of metadata can also be recognized in the above presented metadata structure. In this model, four metadata components can be distinguished: metadata on the Datawarehouse in general, the datasets that are incorporated into the Datawarehouse and the variables and classifications used in the different datasets. These four metadata components are linked to each other which makes it easy to navigate from one component to another.

24. For each component we distinguished several metadata items, taking into account the international expertise on this subject. The metadata items for the variables are listed in the figure below. As an example, we included the documentation on the variable 'removal date' from the dataset of the National Institute for the Social Security of the Self-Employed.

Figure 2. Metadata for the variable 'removal date'

METADATA ITEMS VARIABLES	
Name	Removal date
Abbreviation	Schaansc
Theme	Self-employment
Definition/description	The date on which the association of a self-employed with the National Institute for the Social Security of the Self-Employed is ended.
Derived variable? (+ link to variables from which is it derived)	No
Source (+ link to metadata dataset)	National Institute for the Social Security of the Self-Employed (NISSE)
Validity	1997-now
Measuring level (+ link to codelist)	Nominal (dd.mm.yy)
Specifications/remarks	The deletion date is not always registered on time. As a consequence a number of persons are wrongly registered as self employed. Yearly this affects 15 000 persons, or 2% of the total population of self-employed.
Date last update	8.12.2006

D. BENEFITS OF METADATA

25. When discussing the role of metadata, it is important to focus on the benefits of metadata. To illustrate the importance of metadata in the use of administrative data, some examples are elaborated. First of all, metadata are needed to explain the meaning of variables.

26. Variables on wage generally need accurate documentation, considering the complexity of the notion of income. The European Structure of Earnings Survey (Mittag, 2005) for example uses monthly earnings that are restricted to gross earnings which are paid in each pay period. Gross earnings cover remuneration in cash paid directly by the employer, before deductions of tax and social security contributions. They cover full-time employees as well as part-time employees. The data for part-time employees are grossed-up to those for full-time employees.

27. The Datawarehouse contains a variable 'average daily wage'. The National Office for Social Security calculates this variable by adding up the **normal salary** and the **fixed salary** and dividing it by the number of normally paid days.[6]

28. The **normal salary** is the gross salary before deduction of tax. Rupture fees, bonuses or wages for time spent waiting are not included. What is included then? The salary for days worked, the wage for legal holidays or paid days of absence, the guaranteed wage for incapacity of work, the single holiday allowance (with exceptions), fees granted by the Social Security Fund and fees granted at a closing of a company. The double holiday allowance is not considered as a part of the normal salary. The **fixed salary** is determined by a Ministerial order or a Royal decree. This salary applies to a select group of employees, for example for employees working in offshore fishing or for professional cyclists.

29. It is clear that wage is a complex variable and often country specific. For (inter)national wage comparisons, it is important to understand the components that are included in the wage definition and the way it is calculated.

30. For some variables, it is not only necessary to understand the meaning, but also to know the problems that researchers face when using them. For example, the National Institute for the Social Security of the Self-Employed (NISSE) supplies a variable 'removal date' (see figure 2). This variable indicates the date on which one is no longer inscribed as self-employed. The metadata adds an important remark concerning this variable. The removal date is not always known on time in the database of the NISSE. This means that a

[6] For part-time workers the wage is made equivalent to the wage of full-time workers.

number of persons are wrongfully known as self-employed during the whole year. Only when a new year is integrated into the Datawarehouse, the data are cleaned. On a yearly basis, it concerns about 15 000 persons, approximately 2% of the self-employed population, who might wrongly be classified as being self-employed.

31. Furthermore, the Datawarehouse also contains derived variables. Researchers who use these variables should know how they are calculated. As said before, administrative data have been gathered for administrative purposes and are not always immediately ready to use in socio-economic research. Unlike surveys, administrative data are not built on specifically chosen questions that measure socio-economic indicators. An example is ‘job mobility’. In a survey it would be fairly simple to ask the respondent whether he or she changed jobs (once or more than once) in the last year. When measuring job mobility by means of the administrative data of the Datawarehouse, we are confronted with a few difficulties. These difficulties are also described in the metadata.

32. The statistical unit of the Datawarehouse LM & SP is the 'person' with a unique national insurance number (INSZ). On wage-earners, we have information relating to their employer(s) (i.e. the company registration number). In the Datawarehouse, an employee's ‘job’ is defined as a combination of an employee (INSZ number) and an employer (company registration number). Someone is considered to be job mobile when he or she changes jobs. But since we define a job on the basis of the combination of employee-employer, job mobility is not interpreted as a change of job, but as a change of employer. If the employee is linked to another employer, we speak of job mobility. This means that an employee who changes to a different job but who remains with the same employer is not rated as job mobile. The Datawarehouse only enables us to measure changes of employer.

33. Measuring job mobility faces inevitably the problem of ‘administrative mobility’. This occurs when a company registration number changes for administrative or economic reasons (for example, following a merger or take-over). Without employees changing jobs, their unique relationship employee (INSZ) - employer (company registration number) is altered with the result that they are considered to be job mobile. Changes of this nature occur *collectively* for all employees of the employer concerned and are not considered as job mobility in the Datawarehouse.

E. IMPLEMENTATION OF METADATA STRATEGY

34. The developed metadata model needs to be put into practice. Therefore the administrative sources are systematically screened and metadata are formulated. On a practical level, the documentation process involves learning how each data item was defined by the administrative source, compiling reference catalogues, reading systems manuals and code books, studying the processes behind the data and testing the datasets.

35. Extensive national and international cooperation is needed in the metadata area. On the national level, close cooperation with the providers and users of the administrative data is indispensable to carry out this research project. Because the providers know the strengths and weaknesses of the datasets, the history of variables and the limitations of their administrative data, they can generate important input for the development of metadata. Also the users of the administrative sources are closely involved with the development of the metadata. On a regular basis they are invited for a meeting where the researchers present the progress of the research project and where the users have an opportunity to ask questions, share experiences or give feedback. By using the datasets and analysing the administrative data, users can identify problems or characteristics that need documentation. Feedback on weaknesses found in the data can also be of value to the provider and can lead to a strengthening of the administrative source.

36. In addition also international cooperation and exchange of best practice is necessary. Because there exists a lot of international expertise in the area of metadata, international cooperation and exchange is an explicit aim of this research project. An international network can be of great value to share experiences and compare results.

CONCLUSION

37. The main purpose of the Datawarehouse Labour Market and Social Protection is to collect and store the administrative data sources, held by the various Belgian social security administrations. In order to improve quantitative reporting, policy analysis and socio-economic research, the scope of the Datawarehouse has to be expanded by opening up new datasources or by expanding the existing ones.

38. Within the framework of the expansion of the Datawarehouse Labour Market and Social Protection, our current research project has the intention to make a thorough analysis concerning the content of the data sources, to develop comprehensive metadata and to evaluate the Datawarehouse in view of its international compatibility. The main objective is to achieve a consolidated, harmonised and high quality Datawarehouse

Labour Market and Social Protection.

39. Administrative data are a major resource to provide the government, the public and researchers with a wealth of information and can provide substantial advantages regarding costs and respondent burden. However, using administrative data demands a good knowledge in terms of coverage, content, concepts and definitions, timeliness, frequency and quality. Metadata are indispensable for the correct use and interpretation of administrative data.

40. Of course metadata can never be perfect. They cannot ensure that different users will interpret the administrative data in the 'correct' way. Metadata can only improve and facilitate the understanding and use of the administrative data (Sundgren, 2003, p.24).

REFERENCES

Lindblom, H., & Sundgren, B. (2004). *The metadata system at statistics Sweden in an international perspective.* Statistics Sweden.

Mittag, H.-J. (2005). *European Structure of Earnings Survey (SES).* Statistics in focus. Population and social conditions. 12/2005. European Communities.

Nijs, K. & Berghman, J. (2005). *Validation of the individual part of the social security concept matrix.* Leuven: K.U. Leuven, Research report of the Centre for Sociological Research.

OECD (2006). Data and metadata reporting and presentation handbook. Paris:OECD.

Sundgren, B. (2003). *Developing and implementing statistical metadata systems.* Edinburgh: EPROS.

Vermandere, C., Vanheerswynghels, A. & Van der Hallen, P. (2006). *Eén plus één is drie. Datawarehouse Arbeidsmarkt en Sociale Bescherming: verleden, heden en toekomst.(One and one equals three. Datawarehouse Labour Market and Social Protection: past, present and future).* Gent: Academia Press.