WP.33
ENGLISH ONLY

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

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS)
(Geneva, 9-11 February 2004)

Topic (iv): Using metadata for searching and finding statistical data in websites and portals

USE OF STANDARDIZED METADATA TO FIND, SELECT AND ACCESS STATISTICAL DATA

Contributed Paper

Submitted by Statistics Canada[1]

I. INTRODUCTION

Statistics Canada (STC) is in its fifth year of a multi-year project to implement a central metadata repository in support of its on-line data dissemination activities. The Integrated Metadatabase (IMDB), being implemented in phases, stores information on each of STC's nearly 400 active surveys. The system was implemented in November 2000 and improvements have been made gradually since then. Another milestone was attained in the fall of 2002 with the implementation of hyperlinks from data release articles in The Daily on STC's web site to the IMDB record(s) describing the survey(s) from which the data are produced. The information presently available in the IMDB consists of a general description of the survey, its target population, its methodology - segmented into 10 components - and of measures of the accuracy of the statistics produced. The present development phase aims at storing in the IMDB the names and definitions of the statistical variables produced by each survey, along with the classifications used. At first, only variables disseminated through CANSIM, STC's corporate data dissemination database, will be covered. Users with be able to search and find variables and their definitions in the IMDB, and from there, link to the statistical tables where they are found in CANSIM. This paper reports on the approach adopted to standardize the naming of the variables, list them in various ways on our website and provide access to their numerical values as well as to their definition and classifications.

[1] Prepared by Paul Johanis (paul.johanis@statcan.ca) and Pierre-Paul Bellerose (pierre-paul.bellerose@statcan.ca).

II. FINDING, SELECTING AND ACCESSING STATISTICAL DATA – WHAT'S THE PROBLEM?

STC's on-line data dissemination activities have grown considerably over the years. There are many information modules accessible on its web site containing statistical data. The main ones include: 1) *The Daily*, 2) CANSIM, 3) Canadian Statistics – close to 400 free data tables on many aspects of Canada's economy, land, people and government, and 4) On-line catalogue of STC's products and services. The Census of Population also has an extensive web presence. Finding specific data in these various on-line sources can pose a significant challenge for users, however.

The search engine for the website has been improved over time, by introducing standard metatags on web pages, standardizing the naming convention for web page titles and using a single list of key words and themes for index terms. As a result of these improvements, search results are much more efficient, bringing users closer to the data they are seeking. Nevertheless, the search still produces a fairly approximate result, with users having to scroll through and open many links before truly finding what they want, if at all. For example, the search term "smoking rates youth" returns 6236 hits, with a teacher's kit, **Teacher's Kit: Youth Smoking in Canada: Canadian social trends**, ranked second. A good result, but there are many other hits, including Daily articles and a variety of publications and other products that the user will need to look through.

A second approach has been introduced, which consists of listing Canadian Statistics tables and CANSIM tables by subject, using the standard list of themes. For example, under the theme Health, sub-theme Determinants, one can find a table entitled "Smokers, by sex" in the Canadian Statistics module. In the CANSIM module, clicking through the theme Health, sub-theme Health status indicators, brings up 308 CANSIM tables, of which 14 are concerned with smoking. (To find those quickly from among the 308 tables, the user can search the word "Smoking" using the Find feature in the Windows Edit menu). Again, this brings users fairly close to the data they are seeking, but requires browsing through lists of tables in different parts of the site. In addition, the title of the tables may not tell the user whether it contains a specific variable and it would therefore be necessary to open each table to check.

Statistics Canada has a set of policies requiring a) that any statistical data to be made public must be stored and made accessible in CANSIM before it is disseminated in any form, b) that the data release must first be announced in *The Daily*, and c) that the CANSIM tables and *The Daily* articles offer a link to the related IMDB survey records describing the survey methodology, defining the variables estimated and informing on their accuracy. In addition, the Canadian Statistics tables and the On-line catalogue product descriptions provide links to the related IMDB survey records. As a result of these linkages, users finding a relevant product in any one module can easily and quickly find related products and associated metadata in the other modules of the website, a feature that is described in Louis Boucher's invited paper to this METIS meeting.

III. USING METADATA TO FIND, SELECT AND ACCESS DATA

All of these search methods are concerned with rather large objects, such tables, publications or whole surveys. None of them deal directly with what users are likely to be seeking, that is, data on specific variables. This is what the current development phase of the IMDB can address. To complement the existing search mechanisms on the STC website, the IMDB will provide lists of variables that can be filtered and browsed on the website, access to their definitions and classifications to assist the users in selecting those that meet their needs, and direct links to the CANSIM tables that contain the selected variables. Specifically, these lists will be provided in the following locations:

i) the complete set of statistical variables will be provided in the Definitions, data sources and methods module;
ii) the set of variables produced by a given survey will be provided in the Survey description record, which can be generated from the IMDB from any of the information modules on the website; and,

iii) the set of variables contained in a CANSIM table will be provided with the CANSIM table in the CANSIM module.

The essential challenge in the implementation of this approach arises from the sheer number of specific statistical variables for which data are published - their numbers are counted in thousands - and from the fact that variables that are shared by several surveys are named (labeled) differently in the context of the individual surveys. Allowing users to efficiently find the variables they are interested in requires, first, standardization of the name of the variables, and second, organizing these names into groups with which users may intuitively associate their variables of interest.

IV. NAMING AND DEFINING THE VARIABLES

A. Standardization of the naming of the variables

For standardizing the naming of the variables, STC has complied with the ISO 11179 standard. Using this model will allow variables to be not only comparable across divisions and departments, but also eventually across countries. The basic components of that standard have been retained: data element concept, object class, property, data element and value domain. However, terminology that is more familiar to statistical users is used. The data element concept is simply known as a concept in Statistics Canada, the object class is known as a statistical unit, the data element is a variable and the value domain of a variable corresponds to units of measure or classifications depending if the set of values that the variable can assume is continuous or not.

Following the ISO 11179 conceptualization, a variable (data element) is comprised of a statistical unit (object class), a property and a representation. A statistical unit is an agent (e.g. person), an event (e.g. birth) or an item (e.g. product) about which data are produced. A property is the characteristic of the statistical unit being measured. The representation is the form given to the resulting data, e.g. quantity, value, type, category. All three elements are used to create the name of the variables, e.g. value of sales of an establishment, type of occupation of a person, name of geographic location of a person.

B. Structuring the definitions of the variables

The definition of a variable is provided by the joined definitions of its constituting components, i.e., the definition of the statistical unit involved and of the property considered. The complete definition of a variable also requires the specification of the unit of measure or the classification used to depict the variable, i.e., the non-enumerated or enumerated *value domain* of the variable as it is known in the model.

The non-enumerated value domains correspond to continuous variables, i.e., variables that can assume any value in a set of rational numbers. Examples of such variables are: quantity of product supplied of establishment, value of earnings of employed persons, duration of life expectancy of person. The sets of rational numbers are organized into *units of measures*, e.g. volume in cubic meters, Canadian Dollars, counts in years. In most cases, the same continuous variable may be expressed in terms of many units of measure. Considering the variable "quantity of product exported of establishment", for example, the many possible units of measure include: weight in US tons, weight in metric tonnes, weight in pounds and weight in kilograms. In the case of statistical data disseminated through CANSIM in the form of continuous variables, data users need to know the unit of measure used to quantify the variable. In IMDB, the name of every unit of measure associated with CANSIM variables will be stored.

The enumerated value domains correspond to categorical variables, for example: category of sex of person, name of geographic location of household, or type of industry of establishment. Often, data that are collected as continuous variables are converted to categorical variables for dissemination purposes, for example: category of age of person, or range of time loss of employee. In statistical agencies, the sets of such categories

are organized into *classifications*, standard or otherwise. Classifications are arrangements of classes of descriptors that are mutually exclusive, exhaustive in the universe of interest, and can be aggregated into successive levels in a hierarchy. In most cases, many classifications may be formed for the same categorical variable. The variable "type of industry of establishment", for example, has many classifications, including ISIC, NACE and NAICS. In the case of statistical data disseminated according to given classifications, data users need to know the meaning of each category or class within a classification and the hierarchical relationship between them. In IMDB, every classification will be named, the code and name of each of its constituent classes will be stored and the level and position of each in a hierarchy will be described.

Example of a variable definition display:

Type of Crop Produced of Farm Operation

Type of Crop Produced refers to the reporting of Crop Produced with a classification that describes Type.

Crop produced refers to the type of crop under production by the farm operation.

Farm Operation refers to an operation producing agricultural products with the intent to sell them. Include unincorporated farms with gross operating revenue of $10,000 or more, and incorporated farms with sales of $25,000 or more and for which 51% or more of their sales come from agricultural activities. (Since 1993, farm operations have also included communal farming operations that reported gross operating revenues of $10,000 or more.)

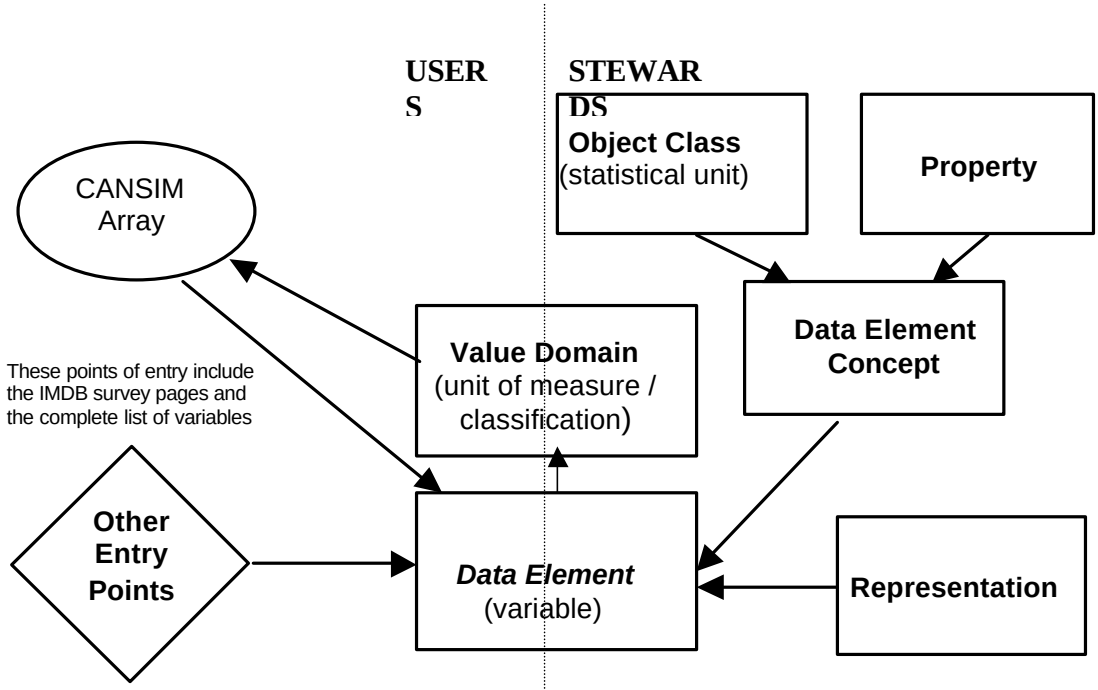
The following classification/s are/is used in reporting statistics:

Crops Categories (click to access the classification description)

C. IMDB statistical variable model

Having described the model used to standardize the naming of the variables and to structure their complete definition, the challenge framework of the present IMDB development phase can be graphically illustrated as follows:

IMDB Statistical Variable Model



Determining the number of statistical units to be defined in the model involves a trade-off. Agents, events and items constituting statistical units can be aggregated to a larger or lesser extent. For example, the fundamental statistical unit "person" can be qualified by adding the property female to make the object class "female persons". More properties can be added to form ever more precise object classes such as employed female persons age 18-25. At the extreme, all properties can be subsumed in the object class so that only the property "occurrence" is left and all value domains become the binomial Yes or No. This would lead to a very large number of variables and complicate searching. On the other hand, if only fundamental statistical units are defined, the more complex properties and value domains become. The compromise adopted in the IMDB model so far consists in defining 75 statistical units about which the 1000+ CANSIM tables display data. In combination with 345 properties and 23 representation classes, the model covers all variables published in CANSIM. Finally, there are 800 enumerated value domains, i.e. classifications, each comprised of varying numbers of levels and classes. These are all manageable numbers, which facilitates search.

V. FINDING VARIABLES AND ACCESSING THEIR NUMERICAL VALUES

The number of variables resulting from the combination of statistical units, properties and representation classes that are disseminated through CANSIM, amounts to about 800. The IMDB model refers to these as *data elements*. When these variables are further specified by the classification or unit of measure used to express their numerical values, the IMDB model refers to them as *specific data elements*. Given that most variables are expressed in CANSIM in terms of numerous classifications, the number of *specific data elements* amounts to over 7000. For example, the variable “category of age of person” appears in CANSIM according to 12 different classifications of age groupings.

The level for which a search function is being developed, in order to access related numerical values in CANSIM, is the *data element* level.

In the IMDB, each variable is linked to the CANSIM tables that display it. Consequently, once a user finds and selects a variable name, he will be presented, as shown in the example in IV.B above, with the variable definition and the list of classifications used in reporting statistics for that variable; in addition, if the user is querying the system from an entry point other than CANSIM, he/she will be presented with the list of CANSIM

table numbers that display the selected variable. Selecting a CANSIM table number, the user will be lead to the data for that variable.

A. The *browsing* approach to finding statistical data

As mentioned in section II above, users will be able to access information on variables from three points of entry: i) from the complete list of variable, by selecting the sub-module "Statistical units, concepts and variables" within the "Definitions, data sources and methods" module on STC web home page; ii) from individual survey descriptions, by selecting the sub-module "Survey information" within the "Definitions, data sources and methods" module; and iii) from individual CANSIM tables accessible from STC web site.

A.1 Alphabetic listing of variables

Conventionally, name lists are ordered alphabetically. The name of the variables being made up of three components, a decision had to be made concerning the one on which to base the alphabetic order. It was decided to base the alphabetic list of variables on the property component, and to list the variables according to the following construction: "property of statistical unit - representation class".

A.2 Accessing variable definitions and their numerical values from the complete list

Data users looking for information related to a given variable cannot be expected to browse through 800 variable names. The entire set of variable names has to be organized in groups and sub-groups through which the users will drill down. The IMDB system offers three browsing options of the whole list of variables: browsing by variable topic, by statistical unit involved in the variables, and by classification domain used in reporting variable statistics.

A.2.1 Browsing by variable topic

For the IMDB purposes, a list of 20 topics and 156 sub-topics has been established. The starting point for establishing that list was the list of subjects established by STC Library for cataloguing STC publications. Subjects that were not pertinent for grouping variables were omitted, e.g. "Reference", or "Statistical Methods", and others were added, e.g. "Industry", or "Supply and Disposition of Goods and Services". See Appendix 1 for the complete list of variable topics and sub-topics.

The process of assigning topics and sub-topics to the *data elements* required considering the associated classification(s) and, at times, the related data file(s). For example, "Type of Disposition of Establishment" is associated with classifications ranging from "Categories of Milk" to "Categories of Coal and Coke". In this case, the variable will have been assigned to the "Agricultural Products" sub-topic of "Agriculture", as well as to the "Primary Industries" sub-topic of "Industry". It follows from this that most data elements are assigned to more than one topic or sub-topic.

A.2.2 Browsing by statistical unit

Often, the subject of research of a user will be a statistical unit, e.g., immigrants, or smokers. The IMDB system offers these users the ability to browse the list of statistical units in order to find and select variables and access their numerical values. For this purpose, we have grouped the units under three macro statistical units: People, Economy, and The State. Each macro unit is further broken down into three possible types of units: agents, events, or others. The agent type is in turn broken down into sub-groups corresponding to *subsets* of agents, *roles* of agents, or *aggregations* of agents. See Appendix 2 for the complete list of statistical units.

A.2.3 Browsing by classification domain

For every variable depicted in CANSIM in terms of one or more classifications, the name(s) of the classification(s) is (are) known. For example, the variable "Type of disposition of farm operation" is displayed in CANSIM according to the following six different classifications: "Agriculture categories", "Harvest categories", "Hogs/Sheep/Lambs categories", "Milk categories", "Vegetable categories", and "Egg categories". The classification name may be a useful starting point for users looking for information related to a research subject. For example, a researcher studying the milk industry could find useful to query the "Milk category" classification to access CANSIM data related to his/her research subject.

The IMDB therefore provides lists of classifications, grouped under the 20 topics listed in Appendix 1. Clicking on a topic name, the user will be shown a list of classifications. Selecting one of them, the user will be presented with the list of CANSIM tables and the list of variables using the classification.

A.3 Accessing variable definitions and their numerical values from the survey web pages

One of the sections of the survey descriptions is titled "Concepts and variables measured". That section will list the names of the variables produced by the survey and disseminated through CANSIM. The number of such variables is never so large as to require grouping them; the variables will simply be listed alphabetically. Selecting any of them will lead to its definition.

Where one or more classifications are used to report the variable, only the name of the classification(s) used in that specific survey will be listed and each classification name will be accompanied by the associated CANSIM table number(s). Selecting a CANSIM table number will lead to the statistics.

A.4 Accessing variable definitions from CANSIM tables

Once users have selected a CANSIM table on the STC web site, they are provided with a grid for selecting data items they wish to extract. From this point, they can click on the survey number(s), which will generate the survey description from the IMDB. In addition, a new button labeled "View variable definitions" will be added below the link to the survey(s) to allow direct access to the related variable definitions.

B. The *search engine* approach to finding statistical data

The browsing approach requires that the user have some idea of the topic, statistical unit or classification domain to which a variable may belong. As an alternative to that approach, the user will be offered the possibility of entering a word, or a string of words, related to his or her interest. A search engine would then search the names of data elements in the IMDB for these words, and associated word(s) in STC Library thesaurus, and return a list of variables that match. From this point, users are then provided with the same options as in the browsing approach described above.

As variables in the IMDB are linked to their classifications, this search mechanism could be extended to individual class names (value meanings) within classifications. Given a word, or string of words, entered in the search engine, the system could first look for synonymous or associated words in the thesaurus, and second, search the class names within each classification for these words. Given the links existing in the IMDB and between the IMDB and CANSIM, the search engine could return lists of *specific data elements* and associated CANSIM table numbers related to the word, or string of words entered in the search engine by the user. Selecting any of the listed CANSIM table numbers would lead users to the data.

Whereas the work is well advanced for the development of the browsing approach, we are only beginning to formulate the mechanics of the search engine. However, the elements needed to offer a very

efficient search mechanism to users exist in the IMDB. The development work on this approach will be tackled in the near future, and its outcome could be the subject of a future presentation.

VI. CONCLUSION

The ISO 11179 conceptualization offers Statistics Canada the possibility of efficiently storing in its corporate metadatabase, and of displaying in a user-friendly way to users, the variable information users need to interpret the statistical data it disseminates. In the longer term, it will also allow the standardization of variables used throughout its statistical programs. As an added bonus, the adoption of the ISO 11179 conceptualization will facilitate interinstitutional and international comparisons.

APPENDIX 1 - List of variable topics and sub-topics

Agriculture
- Agricultural land use
- Agricultural products
- Census of agriculture
- Crops
- Farm finance
- Farmers
- Farms
- Horticulture
- Livestock

Business enterprises
- Business conditions
- Business finance
- Small business
- Type of business

Construction
- Construction materials
- Housing starts and completions
- Non residential construction

Education
- Adult education
- Educational attainment
- Educational institutions
- Educators
- Enrolment
- Fields of study
- Graduates
- Literacy
- Students
- Teaching
- Training

Environment
- Ecology
- Environmental impact
- Environmental Industries
- Natural resources
- Pollution
- Weather conditions
- Wildlife

Government
- Consolidated government
- Federal government
- Government enterprises
- Government finance
- Local government
- Monetary authorities
- Provincial government

Health
- Accidents
- Disabilities
- Diseases
- Health care
- Health status indicators
- Hospitals
- Mental health
- Safety

Industry
- Agriculture
- Arts recreation and Culture
- Communications
- Construction
- Manufacturing
- Primary Industries
- Services
- Transportation and warehousing
- Travel and Tourism

Justice
- Correctional services
- Courts
- Crimes and offences
- Legal aid
- Legislation
- Offenders
- Police services
- Prosecutions
- Victims

Labour
- Employees
- Employers

Employment
Employment benefits
Job search
Labour force characteristics
Labour force participation
Labour mobility
Labour relations
Occupations
Retirement
Salaries and wages
Unemployment
Work arrangements
Work interruptions

National accounts
Balance of payments
Economic conditions
Economic indicators
Financial flows
Gross domestic product
Industry measures and analysis
Input output accounts
Investment and fixed assets
National income and expenditure accounts
Productivity

Personal finance and household finance Consumer spending
Debt
Income
Savings
Social assistance
Sources of income

Population and demography
Age groups
Aging population
Census of population
Citizenship
Ethnic origin
Languages
Migration
Population characteristics
Religion
Vital statistics

Prices and price indexes
Agricultural price indexes
Construction price indexes
Consumer price index
Export price indexes
Import price indexes
Industrial price indexes
Inflation
Price indexes
Prices
Raw materials price indexes

Science and technology
Information technology
Innovation
Intellectual property
Inventions
Sciences
Technology

Supply and Disposition of Products and Services
Disposition
Inputs
Inventories/Stocks
Orders
Production
Revenues/Expenses
Sales/Receipts
Shipments

Social conditions
Child care arrangements
Families
Households
Housing
Social behaviour
Support groups

Target groups
Time use

Trade of Goods and Sevices
Exports
Free trade
Imports
International trade
Interprovincial trade

Transport
Cargo
Freight
Fuel
Infrastructure
Modes of transport
Vehicles

Travel and tourism
Commuting
Domestic travel
International travel
Passengers
Tourism

APPENDIX 2 - List of statistical units

PEOPLE – Agents

Agents by age or sex

- Children
- Person
- Person 12 years or over
- Person over 15 years
- Women

Agents aggregations

- Census family
- Economic family
- Household
- Population

Agents by role and domain

Justice

- Crime victim
- Criminal suspect
- Homicide victim
- Legal aid lawyer
- Person accused
- Person appearing in court
- Person charged
- Police officer
- Police personnel

Citizenship/migration

- Emigrant
- Immigrant
- International migrant
- Interprovincial migrant
- Intraprovincial migrant
- Non-permanent resident

Labour

- Employed person
- Employee
- Employment insurance beneficiary
- Labour force participant
- Paid worker
- Self-employed worker
- Unemployed person

Other

- Household head
- Mother
- Smoker
- Student
- Traveler

PEOPLE - Events

- Birth
- Death
- Divorce
- Homicide
- Marriage

ECONOMY - Agents

- Company
- Economy
- Enterprise
- Establishment
- Farm operation
- Institutional unit

ECONOMY – Others

- Business Location
- Crop
- Dwelling
- Farm input
- Help wanted advertisement
- Job
- Land
- Passenger-Kilometer
- Person-trip
- Person-visit
- Product
- Transaction
- Trip
- Vehicle
- Vehicle-Kilometer
- Want ads

THE STATE

- Building permit
- Charge
- Criminal case
- Claim
- Court
- Criminal incident
- Employment insurance claim
- Legal aid application
- Legal aid plan
- Sentence
- Shelter admission
- Shelter