

18th Census Population Conference
Disseminating Census Data with Beyond 20/20

Bill Lidington

Ivation Datasystems Inc.

Historical Overview

Beyond 20/20 is a software product that was designed and built for statistical data disseminators. The product features are there because of the demands of these disseminators. The evolution of the product continues to be customer-driven – as technology changes and as the markets for statistical data evolve, so will Beyond 20/20.

In 1993, Ivation released Beyond 20/20 for Windows 3.0. In the same year, the Labour Force Survey Division of Statistics Canada purchased licenses of Beyond 20/20 for internal use. By the end of 1994, several subject matter divisions of Statistics Canada were using Beyond 20/20, both for internal analysis and for dissemination on diskette and CD-ROM.

In the middle of 1996, the Census Division evaluated their options and decided to use Beyond 20/20 for most of their digital dissemination on CD-ROM products. Statistics Canada had developed their own product called C91 for dissemination of the 1991 Census results. This product was written in BASIC. With the advent of Windows, and the continuous evolution of technology and operating systems, Statistics Canada decided that it made more sense to buy an existing off-the-shelf product, than to attempt to build and maintain one in-house. Census Division was also being pushed toward adopting Beyond 20/20 by some of the major Census data users who had already adopted Beyond 20/20 as an internal data storage and access tool. These federal agencies wanted their standard and custom Census tabulations to be delivered to them in Beyond 20/20 format.

By this time, the Data Dissemination Division of Statistics Canada had become a center of expertise in the use of Beyond 20/20. They had created digital products based on Beyond 20/20 which were accepted in the marketplace. They operated an internal Help Desk for support of the Beyond 20/20 authoring tool. This infrastructure made it easy for Census Division to make the decision to use Beyond 20/20.

When Statistics Canada decided to use Beyond 20/20 for dissemination of Census data, they created a list of enhancements which Ivation committed to include in its next version. This version, called Version 5.0, is now in Beta release.

Several domestic statistical agencies are now evaluating Beyond 20/20 for dissemination of Census summary tables and public-use microdata files. The following two case studies describe (1) Statistics Canada's use of Beyond 20/20 in Census dissemination and (2) the French INSEE's intentions with respect to Beyond 20/20 and dissemination of their decennial Census results.

1. A Brief Description of the Dissemination System for 1996 Canadian Census

Background

Prior to the 1991 Census, Statistics Canada had used mainframe systems for processing the Census and for building products for dissemination. Statistics Canada employees built almost all of the mainframe software related to the data base management system and the tabulation. Most of the outputs from the mainframe system were on paper, which was becoming expensive. A move to more electronic output had to be made. The facilities for metadata management had to be improved. The dissemination group of Census found the existing system to be too expensive and inflexible for their needs.

A decision was made to build two systems, which would be integrated. The first, now called the Computer-Assisted Product Specification System (CAPSS) was to be the main tabulation engine. It would be used to build aggregated data tables which would become part of either standard data products or custom products built on request.

The second system is called the Electronic Shelf. Its main purpose would be to hold and maintain a table repository for tables built by the CAPSS system. This system would be accessible by disseminators of Census data within Statistics Canada.

A. The Computer-Assisted Product Specification System (CAPSS)

Development of the new system began in 1989, in preparation for dissemination of the 1991 Census results. A decision was made to use SYBASE on Silicon Graphics Inc. UNIX computers. The new system would use standard relational technology in conjunction with TPL tabulation software.

Special attention was given to the management of metadata at all levels of the data. Documentation and definitions related to variables were stored in both official languages.

Tabulation rules related to the use of specific variables were stored related to each variable. In this way, the system became an intelligent system, without the users having to know or remember rules related to each variable each time a variable was used in a tabulation.

A Graphical User Interface (GUI) was built so that tabulations could be designed by almost casual users without them having to learn a language of any sort. The tabulation specifications were then compiled into script files that the tabulation engine could understand.

Table specifications are stored in the CAPPs system so that they can be reused. Clusters of table specifications can be stored, so that all tables related to a single product can be retabulated automatically and output to a single directory. Access to the SYBASE data base was optimized by writing native C language access to the relational files. This had a dramatic effect on tabulation from the large micro-data files related to the Census. Records were accessed from the data base and streamed to the tabulation engine. Following tabulation, the Statistics Canada confidentiality rules and rounding rules are applied to the data in the table to ensure confidentiality is preserved. During this process, data points are suppressed or rounded according to a documented set of rules.

The format of outputs from CAPPs are specified by the user through the GUI. The following output formats are supported: paper, WKS (lotus spreadsheet), fixed-formatted ASCII, BEYOND 20/20, and Autoload to Electronic Shelf.

Creating Tables in BEYOND 20/20 Format

Through the user interface, a person making a tabulation request can specify that the output format of the table be in BEYOND 20/20 format. This is accomplished through the use of the BEYOND 20/20 INTEGRATOR'S BUILDER which runs on the Silicon Graphics Inc. UNIX environment. Following the creation of the tabulation, source ASCII files describing both the data and the metadata are dynamically prepared according specifications provided by Ivation. The BEYOND 20/20 INTEGRATOR'S BUILDER is then dynamically called to convert the tabulation into BEYOND 20/20 format. Tables that are output to BEYOND 20/20 format are completely bilingual, with extensive metadata at all levels of the table. Table titles, table descriptions, variable descriptions, codebook labels, and variable-value descriptions are all provided.

B. The Electronic Shelf – the Tabulated Data Warehouse

This system is also SYBASE-based, and it utilizes relational technology. The Electronic Shelf is the tabulation repository, which holds all the standard tables related to a Census. There can be hundreds of tabulations, which comprise standard and non-standard data products.

Who Uses the Electronic Shelf?

The Electronic Shelf is an in-house system accessible only by Statistics Canada employees. The users of the Shelf are all those people in Statistics Canada who are involved in the dissemination process. These are people at headquarters who are involved in creating CD-ROM-based data products. In addition, the regional offices of Statistics Canada each have Census data dissemination experts. These people sell the standard data products in addition to responding to ad hoc requests for Census data.

How Does the Tabulated Data get into the Electronic Shelf?

All tabulated data originated on the CAPPs system. One of the output options in creating tables on the CAPPs system is 'AUTOLOAD TO SHELF'. Choosing this option creates a temporary file that can be directly loaded into the Electronic Shelf system.

In actual practice, users of the CAPPs system find that tabulations created there usually are output into temporary files where they are checked and modified manually prior to loading into the Electronic Shelf.

Accessing the Electronic Shelf

Users access the Electronic Shelf through a Graphical User Interface. Tables stored on the shelf can be at various stages of release, from preliminary to final. Access to preliminary data is restricted and controlled using passwords and the controlling features of the SYBASE software.

Since there can be hundreds of tables on the shelf, the users first need to find the table they are looking for. An extensive, indexed, metadata searching facility is available so users can find the tables they want to extract. Users can search either by keyword or by subject-matter category in either official language.

Once a table is identified as being the correct one, the user can then subset the table by selecting geographies for inclusion into the extracted table. A comprehensive system for selecting geographies has been implemented to satisfy all the needs for selection. For example, alias place names can be used during this process.

Then, the user can subset the variables (or dimensions) to be included in the extracted table.

Various output formats are available for extracting data from the Electronic Shelf. BEYOND 20/20, comma-delimited, fixed formatted ASCII, and C91 (a Statistics Canada 1991 Census dissemination format) are supported.

Regional Dissemination using BEYOND 20/20 Format

Census dissemination personnel in the regions have now become used to responding to ad hoc requests for information. They frequently answer these requests by sending out tables in BEYOND 20/20 format. This is how the process works.

Regional employees of Statistics Canada access the Electronic Shelf to find the tables that will be used to respond to the request for information. They choose BEYOND 20/20 format and download the tables to a single directory on their personal computers.

They then use a custom-made product called the PACKAGER which creates a new directory including the BEYOND 20/20 BROWSER, a setup routine, all the relevant meta-data files, the license agreements, and the data tables. Electronic Quickstart Guides showing how to use BEYOND 20/20 BROWSER are also included.

Optionally, they can include diskettes containing the BEYOND 20/20 BROWSER User Guide in Adobe PDF format.

They then choose the media to ship this custom product to the customer. They can choose between diskette and CD-ROM.

How Does a Data Customer Get HELP If It Is Needed?

The regional representatives act as the first line of support for the data customers. They respond to the data customers' needs for support in loading and accessing data.

If the regional representatives are unable to answer a question, they call the central Census HELP DESK to find answers to questions they have about either the data or the software accompanying the data.

Standard Data Products in BEYOND 20/20

Many standard data products related to the 1996 Census were published on CD-ROM. The dissemination group of Census Operations Division used a combination of two software products – ICON AUTHOR and BEYOND 20/20 to create integrated data products for public consumption.

The tabulations are created on the CAPPs system. They are then checked and edited manually. The metadata-to-data linking and integration work is performed by Census programmers, who create the final, bilingual product for public consumption. A Master CD-ROM is created in Census Division and sent over to Dissemination Division for replication and publishing. After replication, these products are catalogued and sent out to the regional offices for distribution to data customers.

An example of a standard data product from the 1996 Census is the '**Nation Series**'. 'For years, federal, provincial and municipal governments have used the **Nation Series** as a benchmarking tool to compare trends in different parts of the country. The 1996 Census Nation data are hyperlinked to metadata and bundled with user-friendly software, BEYOND 20/20™ for Windows™.' (March 1988, *Information Matters*, Statistics Canada, page 4).

Later in 1998, the Dimension Series of products and Public User Microdata Files (PUMF) files from the 1996 Census will be released in BEYOND 20/20 format on CD-ROM.

Dissemination Plans at l'INSEE for the Decennial Census

Background

The data collection for the decennial census in France will occur in 1999. L'INSEE will collect detailed information from just under sixty million people who live in France. Collection of this information is a legal requirement according to the laws of France. The information collected is used for a variety of legal purposes, from the distribution of federal grant programs to the regions to the salaries of the mayors of every municipality.

In France, unlike Canada and the United States, there is only one Census form used as the data collection instrument. These forms are filled out at the household level, and scanned into digital form for further processing. Over the years, France has developed an algorithmic methodology for the edit and imputation process. This has resulted in the development of a computer program which, for the Census, contains a very large number of algorithms which are used to cleanse the data and impute missing responses. L'INSEE has a legal requirement to reduce the error rate in Census results to less than one-sixth of one percent at the lowest level of detail.

Dissemination Plans

The Dissemination Problem

France is divided into twenty-four statistical regions. Each INSEE regional office has the responsibility of serving the information needs of organizations and individuals in each respective region. The information requests subjected to each regional office are diverse in nature. Some data-related questions can be answered over the telephone. Other enquiries are complex and require written specifications. The people making the requests are the full range of data customers, from students writing research papers, to large companies linking their sales data to demographic data, to international organizations and other countries.

The Solution

The approach l'INSEE has adopted to prepare for dissemination is a comprehensive one. In the year 2000, the first standard products related to the decennial census will be distributed. In France, the measures first released are called the 'Blue Book' and these variables would roughly correspond to those collected in the Census Short Form of Canada and the United States. In 2001, the detailed Census data will be released, containing the full range of variables resulting from the detailed Census form.

The statistical offices of federal agencies will have privileged access to census data including online access to data residing on INSEE's mainframe computer. Within INSEE, each of the regional offices will be equipped with the Beyond 20/20 Builder, the tool that is used to author Beyond 20/20 tables and micro-data files. Each region will be trained in the use of the Beyond 20/20 Builder, and will be proficient in the use of the Beyond 20/20 Browser, so that each region will be able to support the customer's use of the distributed data product.

To prepare for the onslaught of information requests, each region is expected to make a set of data tables that will respond to the majority of information requests. These tables will be immediately accessible to those responding to information requests. In addition, it is expected that information officers in the INSEE regional offices will use the Beyond 20/20 tabulation facility to dynamically build tabulations to respond to ad hoc requests for information.

It should be noted that each region also has a number of other software tools available to them for the creation of custom data products. In fact, each region, as an independent body, has the authority to create its own standard data products, if there is a perceived need in the customer base. SAS is one such product that is used heavily in the regional offices for data analysis requirements.

The Dissemination Division of l'INSEE will create standard products on CD-ROM. The products created will contain Census profiles, which are shown in national and regional data tables. The customers of these products will use the Beyond 20/20 Browser to view these tables. Each regional office will be able to sell these CD-ROMs to their data customers. Later, in the dissemination flow of products based on the Census, more detailed data products will be created, and public-use microdata samples will be disseminated.

Alternatively, if the standard data products do not meet the needs of the data customers, each region will be able to dynamically create responses to the data requests – in the format requested by the customer. Beyond 20/20 will be one of the formats available to data customers who make ad hoc requests for information. Each region will also have the capability to disseminate flat ASCII files and spreadsheets, for example, on the media chosen by the customer, (paper, diskette or CD-ROM, for example).

In conclusion, the dissemination strategy that l'INSEE has adopted is a geographically distributed one, and it is a tools-based approach so that all information requests can be met, regardless of data content or data format required by the customer. The emphasis here is on service to the customer.

The Future

It is likely that this will be the last Census taken in France. There is a move to continuous measurement using survey methodology for the same reasons that the Americans are moving in this direction. The reason is simple: Census data is used for a variety of legal purposes, therefore its timeliness and accuracy are too important to be collected once every ten years. Moving to continuous measurement is expected to have an impact on dissemination in that more customers will ask for data more frequently. It will be important to have flexible software tools in the hands of both the disseminators and the data customers.

The French decennial Census will use the Internet for dissemination after the initial release of 2001 data products. The Internet has been slow to gain acceptance in France due primarily to the cost of access. When the Internet becomes a mainstream technology in France, it is expected to have a dramatic effect on the volume of public inquiries and l'INSEE's ability to respond to these inquiries in a timely manner.