Fourth International Conference on Agriculture Statistics
Beijing, China
October 22-24, 2007

How to Build an Integrated Database in Agriculture and Food: The Farm Income and Prices Section Database – A Case Study[1]

Denis Chartrand[2]
Director, Agriculture Division
Statistics Canada
12th Floor Section B-8, Jean Talon Building
170 Tunney's Pasture Driveway
Ottawa (Ontario), Canada
K1A 0T6
Email: denis.chartrand@statcan.ca

August 31, 2007

[1] The opinions expressed in this paper are those of the author and do not necessarily reflect the official position of Statistics Canada.

[2] The author wishes to acknowledge the contributions of Marcelle Dion, Bev Hammond and Paul Murray of Statistics Canada in the preparation of this paper.

Abstract

The Farm Income and Prices Section in the Agriculture Division of Statistics Canada is responsible for a regular program of collecting, compiling, analysing and disseminating aggregate agriculture economic statistics. Data from these series flow to several divisions in the Canadian System of National Accounts Branch to form the agriculture sector's contribution to GDP, as well as to many key users outside Statistics Canada both in government and in the private sector, and together are considered vital indicators of the health of the farm sector. In producing these data sets, the Section integrates a tremendous volume of data on a monthly, quarterly, semi-annual and annual basis from a wide range of sources within the Agriculture Division, other divisions in Statistics Canada, other federal and provincial government departments and agencies, producer marketing boards and industry associations. The rapidly changing structure of the industry, with its escalating complexity and diversity, as well as a loss of critical administrative data sources, have increased the difficulties of data collection and estimation. In an environment of resource constraint, ensuring an efficient production process that leverages advances in technology to permit sufficient time for analysis to ensure the high level of data quality essential in meeting users' requirements is a challenge. This paper provides an overview of the issues, the strategy and the considerations surrounding the development of a new database to be used in the Agriculture Division to facilitate the process.

Introduction

Information about the incomes and economic well being of farmers and their families is of broad interest all over the world. In recent years, low aggregate farm income in Canada has been widely reported by the media and concerns have been expressed about the financial health of the agriculture industry and its participants. As farm operations have become more complex and more structurally diverse over time, both the challenges of producing high-quality data and the need for improved measures and tools for assessing industry performance and farm family financial well-being have increased. This, in turn, has increased the demands on the processing systems and related data storage systems to provide more flexibility to adapt to these requirements.

The traditional aggregate net farm income measures produced by STC are part of a broader set of economic accounts. For all of these accounts, the estimates are intended to represent all primary agricultural activity, i.e., to cover all farms as defined by the Canadian Census of Agriculture. Under the umbrella of its Agriculture Economic Statistics (AES) series, the Farm Income and Prices Section (FIPS) of the Agriculture Division publishes a Farm Product Price Index and agricultural product prices monthly; farm cash receipts data quarterly; and annual series on operating expenses and depreciation, net farm income, income in kind, value of inventory change, program payments, debt outstanding, capital values, cash flows, value added and a balance sheet. All of these data sets are available electronically and are conceptually and operationally integrated into a database to provide a broader and more complete picture of sector performance.

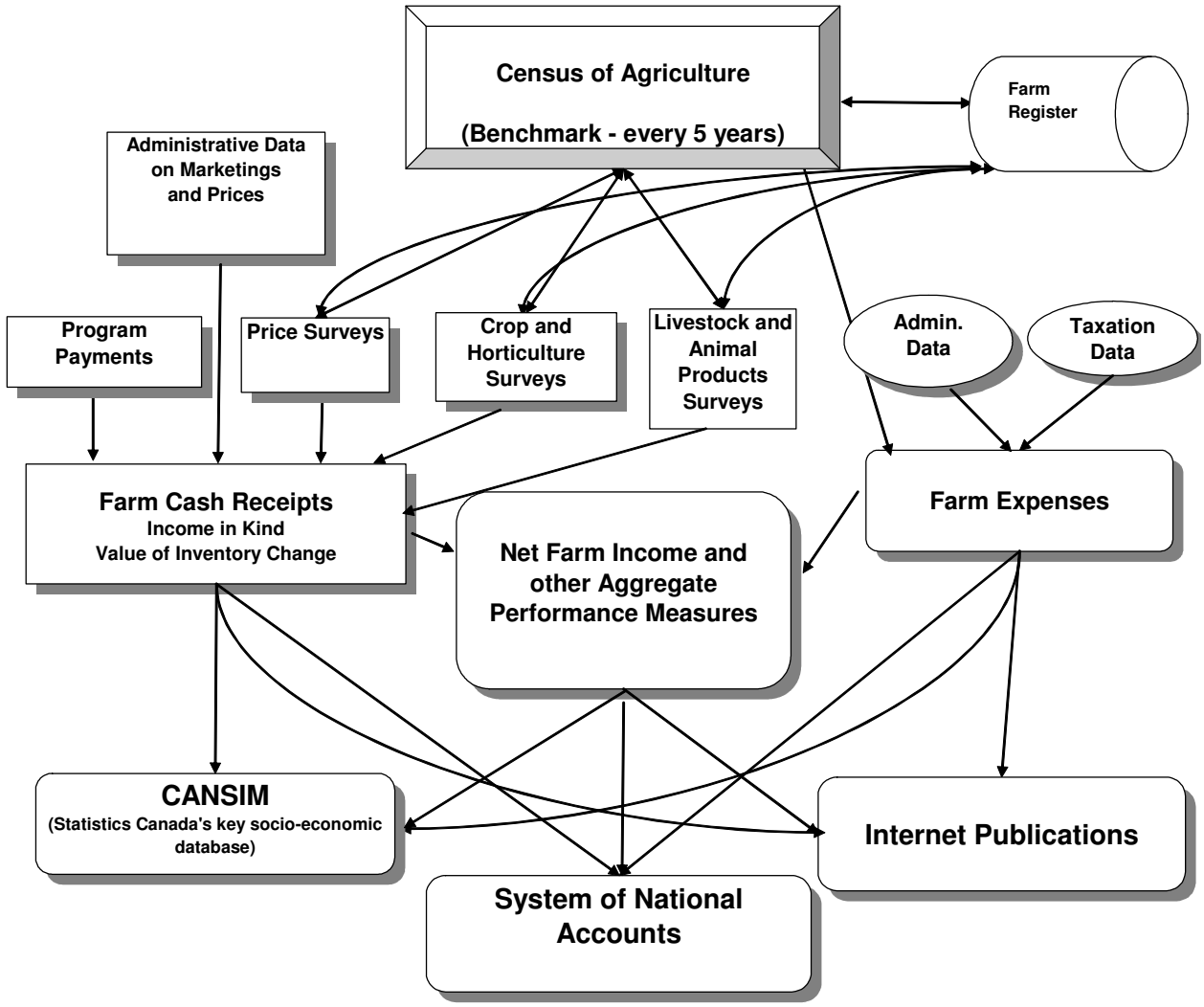
As data from these series flow to several divisions in STC's System of National Accounts (SNA) Branch to form the agriculture sector's contribution to Gross Domestic Product (GDP), they must conform to international standards which have long provided a solid conceptual framework for collecting and compiling aggregate agriculture economic data. These international standards are reviewed on a regular basis — such as revision of the United Nations' System of National Accounts in 1993 (SNA 93) — to ensure their relevance (Caldwell and Murray, 2005).

Over the roughly 80 years of producing farm income estimates, a large variety of survey and administrative data has been used to produce the revenue and expense accounts. In producing the current aggregate agriculture economic statistics series, FIPS integrates a tremendous amount of data from many sources within Statistics Canada into its database. In addition, to reduce respondent burden and cost wherever possible, administrative data from federal and provincial governments and agencies, marketing boards and producer organizations are used in place of farm surveys.

The evolution and complexity of the agriculture industry and of users' requirements for enhanced and better integrated information will be more fully explained later in the paper. Suffice it to say for now that changes in the Canadian agriculture industry have contributed greatly, along with the ongoing loss of key administrative data sources over time, to the complexity of ongoing processes for data collection, compilation, analysis and dissemination faced by FIPS. The need for a highly efficient and effective integrated processing system has become critical to accommodate new data needs for policy development and program monitoring.

As shown in Figure 1, the current collection activities mentioned above are an integral part of the Agriculture Statistics Framework that feed into the SNA for the calculation of the GDP for the agriculture industry.

Figure 1 – Agriculture Economic Statistics Framework



Data quality concerns are considered in light of the six dimensions of data quality contained in STC's Quality Assurance Framework (relevance, accuracy, timeliness, accessibility, interpretability and coherence) (Statistics Canada, 2003). Dion (2007) noted that an integrated statistical framework plays two important roles: a quality assurance role and a "fitness for use" role facilitating the interpretability of the information. The quality assurance process of the agriculture statistics program is carried out in two phases: on an annual basis with the provision of information to SNA, and on a quinquennial basis with the release of Census of Agriculture (CEAG) data. The development of a more integrated database in Agriculture Division will enhance analytical capabilities for better quality assurance.

As shown in Figure 1, the Farm Register provides the frame for most divisional survey activities; the crop, livestock and financial data serve as inputs into the derivation of the farm income series. In turn, the production of the farm income estimates is a coherence exercise that allows the validation and revision (if necessary) of the input data. Every five years, the CEAG allows for an update of the Farm Register used to draw samples for the various crop, livestock and financial surveys and a base to revise estimates (including farm income estimates) produced between censuses. In turn, sample surveys and the farm income series are useful tools to validate census data.

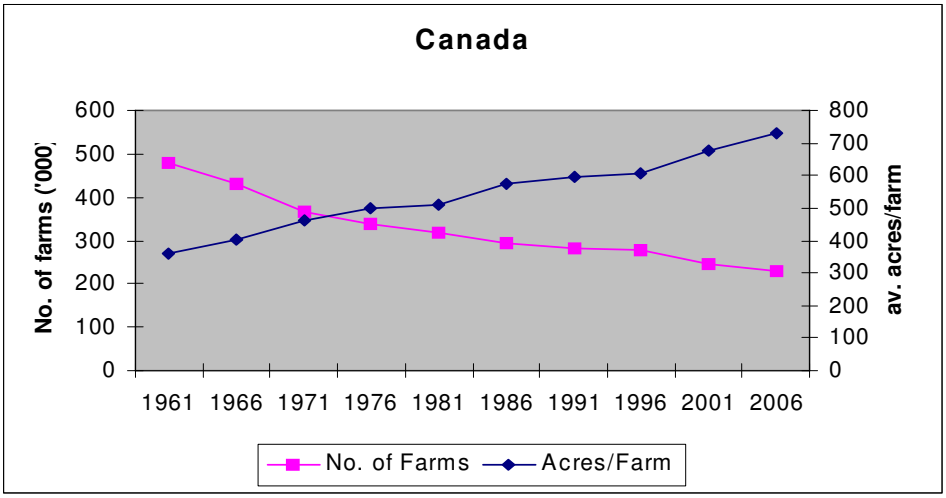
1. Evolution of Farm Structure and Farm Income Issues

The Canadian agriculture and agri-food system is a complex integrated production and distribution chain of industries that supplies food, beverage and tobacco to domestic and international consumers. It is an integral part of the global economy with trade occurring at each stage of the chain (Agriculture and Agri-Food Canada, 2006). However, the relative contribution of primary agriculture to GDP and employment has been declining significantly. Although the value of agricultural production has tripled since 1961, the Canadian economy as a whole has grown at a faster rate (by six times), driven mainly by growth in the high-tech and service sectors. The result has been a drop in the share of primary agriculture to about 1.5% of GDP. On the other hand, the agriculture and agri-food industry as a whole remains a significant contributor to the Canadian economy, accounting for 8.1% of total GDP and 13.1% of employment in 2004.

The rapid pace of the evolution in the structure of the agriculture industry can be partially illustrated by examining the changing number and size of farm operations in Canada. After peaking in 1941, farm numbers have been steadily falling, while at the same time, average farm size has grown with the relative stability of total farm land over time in Canada (Chart 1). Moreover, agriculture production has become much more concentrated on larger farms. Chart 2 shows that an increasingly smaller proportion of farms account for the majority of sales over time.

Chart 1 - Farm numbers and average farm size

The trend towards fewer but larger farms continues in Canada

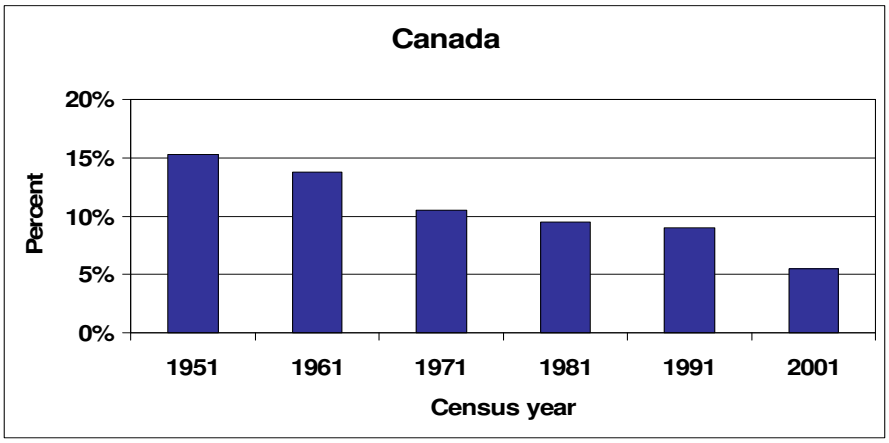


Source: Census of Agriculture

As farms become larger, with more complex legal and operational structures, the tasks of

collecting, processing and analysing data for the industry, and ultimately, measuring its performance, have also become more complicated and difficult. The growing importance of vertically-integrated operations, increased contractual arrangements and more varied marketing opportunities such as direct marketing/dual markets/numerous payment options have added to this complexity. These factors point to the need for a database that will offer flexibility to the analysts in their measurement and interpretation roles.

Chart 2 - Smallest percentage of farms needed to account for half of the agricultural sales
Concentration has been increasing – sales becoming more concentrated in larger operations



Source: Census of Agriculture

2. The Role of the Farm Income and Prices Section (FIPS) in Measuring Farm Income within the Statistical System

2.1 Evolving Needs

Statistics Canada has produced provincial and national estimates of annual farm income and capital value dating back to 1926. FIPS has the responsibility for producing most of Agriculture Division's aggregate economic data. As the structure of agricultural production and marketing began changing more rapidly in the latter part of the twentieth century and policy agendas of governments expanded, users of agriculture financial data identified needs for more detailed and integrated statistical information that will illuminate current issues.

2.2 The Evolution of the Current System for Processing Farm Income and Prices Data

Just as the agriculture industry and the measures of farm performance have evolved over time, so has the collection, compilation, analysis and dissemination of the data. Again, as technology has increased productivity in farming, it has had a tremendous impact on the way Statistics Canada does business. The advent of mainframe, then more and more powerful midrange and micro computers, has transformed the processing of data, while increasingly sophisticated methodologies and collection options led to stratified probabilistic survey samples collected with computer-assisted telephone interviewing (CATI) approaches from regional offices. Respondent burden has been reduced due to the increased use of administrative data and through integrated methodologies and

electronic options for data reporting (including secure internet) as part of Statistics Canada's streamlining initiatives.

The organizational structure of work changed more slowly as the transformation from paper to micro computers increased efficiency but not the nature of the work. Capital replaced labour to some extent in processing but the analysts and their staff adapted to increased technology in steps, often determined by the technical skill of the clerks/analysts of the day. As spreadsheet capacity increased, more complex calculations were automated, based on the ability and time of individual staff, resulting in a patchwork of improvements by one analyst and a completely different approach by another analyst working on another set of commodities.

To improve processing efficiency, FIPS implemented its current Visual FoxPro[3] database in February 1998. While major benefits included the automatic calculation of aggregates by the base eliminating the need for manual entry of data into summary reports, better verification and reporting features, automatic loading of files to SNA and STC's CANSIM base of Canadian socio-economic data, and easy access to data for completing user requests, the highly interconnected spreadsheets continue to pose a challenge to the timely production of accurate data.

With the aim of addressing the issues raised above and explained further in Section 3.2, Agriculture Division embarked on a project to develop a division-wide system to streamline and improve the efficiency for processing aggregate data. As part of the project, FIPS staff began a review of all aspects of their data flows with divisional systems analysts. The goal of the project was to ensure that the vast amount of data that make up the Section's many series are received, processed, analysed and disseminated as efficiently as possible, resulting in the availability of very high quality data for users.

3. The New System For Processing Farm Income and Prices Data

3.1 Challenges posed by the environment

- Due to the demanding nature of the work in FIPS, staff turnover has been high in recent years. FIPS staff work with few resources under constant time pressure, resulting in difficulties in scheduling time for training staff and documenting procedures.
- Due to the complexity of the data, different units within FIPS work with different sets of data and on different timeframes for various data series. Yet each of these groups must be able to access data from the other groups with an understanding of which data are preliminary and which are final.
- STC has an accepted model for data warehousing. One of the biggest questions the systems analysts had to grapple with in this project was: is this a data warehouse[4] or a transactional database[5].

[3] Visual FoxPro: a software for creating databases and user applications (screens).
[4] The STC Data Warehouse Framework was not considered for this project because it did not allow for manual updating of business rules.
[5] Transactional (operational) database: a database where data are being assembled for analysis (integrated, cleaned and corrected, value added). In such a system, users can manually correct data and change their business rules on the fly in order to adapt to changing assumptions.

3.2 Weaknesses of the existing system

The processing of data within Agriculture Division has become a more complex and somewhat an inefficient process over time. As the major Agriculture Division integrator of data from a wide array of survey and administrative sources, nowhere is this more evident than in FIPS where the data are amassed, analysed, integrated and formatted before output to the SNA Branch and outside of STC. Some of the key issues that required attention are listed below:

- A weakness with the current myriad of Microsoft Excel[6] spreadsheets is the lack of a standardized structure that imposes more discipline on how they are used. Having been set up by various people over the years, there is a need for more consistent formatting.
- The business rules are sometimes hidden by the complexity of the links, making it difficult for the analysts to see how the data point is being produced. In addition, spreadsheets can be modified without requiring additional permissions and without any audit trail to inform other users.
- The spreadsheets are used for many different functions, including assembling and capturing data from different sources, calculating new variables, calculating different geographic levels, calculating values for different time periods, producing validation reports and setting confidentiality flags. Because Microsoft Excel is positionally dependent, the same data point may need to be repeated in multiple spreadsheets in order to fulfill all of these functions. Ideally, there is only one true copy of the data point and all others are links, but there is nothing in the system to enforce this.
- The current systems rely on individuals to maintain and process them, often with limited back-up. Whenever staff leave, the learning curve for new employees on the details of each specific spreadsheet relating to their task is much longer than desirable, and they cannot easily carry that expertise over to a new task. Reducing confusion and improving coordination would greatly reduce the risk of error.
- The current Visual FoxPro database used by FIPS is written in software that will be phased out starting in 2009 (although support will continue until 2015). Systems analysts are required to program key reports and resolve more complex production problems.

3.3 Options for an upgraded system

The systems development team, from Agriculture Division considered several design options for the new system:

1. Move the database to SQL/Server[7] but leave the processing in Microsoft Excel.
2. Move the database to SQL/Server and move the processing to a more structured toolset.
3. Move all data and processing to a SQL/Server database.

The major factors to consider when choosing a solution include software and training costs, software acceptance within STC, ease of use and sustainability for clients. Whatever the solution, FIPS staff must be able to define and load their data, understand and change their business rules, and create and run their own reports without reliance on the system developers.

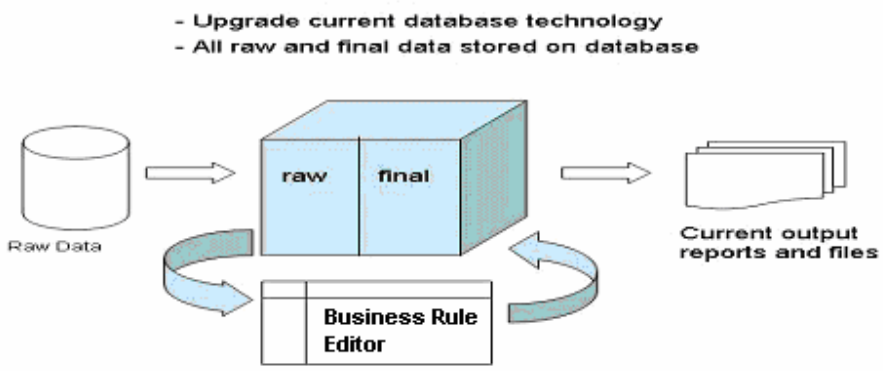
[6] Microsoft Excel: a spreadsheet software.
[7] SQL: Structured Query Language, the industry standard for all databases. SQL Server: a software for creating and managing databases.

After evaluating the benefits of each option and the capabilities of various software (including SAS[8], VB.NET[9], SQL/Server, Microsoft Excel and their business intelligence tools), the team decided on Option 2 which is displayed in Figure 2 below.

Proposed System (Option 2): Move the database to SQL/Server, and move the processing to a more structured toolset (i.e., Separate the load (raw data) and Business Rule functions and upgrade the database)

In this option, subject matter staff (FIPS) would first load all raw data to the database. This would eliminate the need for linked spreadsheets to assemble data and verify the data. They would then use a separate computation tool (Business Rule Editor[10]) to compute new values and load the results to the 'final' database. The possible options for the Business Rule Editor are Microsoft Excel templates, SAS and VB.NET.

Figure 2 - Proposed System



3.4 Project Life Cycle: Analysis and Design Phase

Upon adoption of Option 2 as the design for the system, FIPS began mapping the data flows into and out of each area of the section. It then conducted a pilot project in January 2007 — the FIPS Aggregate Data System Pilot Project. The purpose was to investigate different application models and commercial software, gather user requirements and, using a representative set of data, attempt to understand enough of the business processes to make recommendations on a system design and choice of software.

It quickly became apparent that the bulk of the work was not systems-based at all, but rather to specify the thousands of business rules buried in the existing Microsoft Excel spreadsheets. Although one option would have been simply to tidy up the spreadsheets, the team felt it was important that these business rules be formally logged, documented and reviewed instead of just blindly re-implementing them in the new system. This will require, however, a large resource and time commitment from the Division, probably extending the project well into 2009.

---

[8] SAS: Statistical Analysis Software, a software for creating databases and user applications, especially for statistical purposes.

[9] Visual Basic for the Microsoft.NET environment: a software for creating user applications (screens), often used with a database software such as SQL/Server.

[10] Business Rule Editor: a tool that allows clients to create new data points using calculations (business rules).

The plan is to analyze, design, implement and test the business rules for one commodity or group of commodities at a time, until all of the FIPS business rules are implemented. The development of the system itself will begin in October 2007 and should be completed well before all of the business rules are specified and implemented.

3.5 Design Challenges

- **Disciplined workflow and quality assurance measures**

A change of system is an excellent time to review and update the standards and Quality Assurance/Quality Control (QA/QC) measures implemented in FIPS. Systems development staff need to be aware of these measures so that the new system can be aligned with them where possible.

Standards that affect the system include permissions for: who can change data and formulas, who needs to be informed of what kinds of changes; naming conventions (files, directories, system objects, business rules/formulas); file handling; documentation; and all other processes that affect data quality and process efficiency. Such standards are easier to maintain over time if they are assigned not to a specific person, but to a specific role. Then, as people come and go, they can be easily given a list of their permissions.

- **Flexibility versus structure**

Users are accustomed to a current "system" where they have virtually unlimited flexibility, access and ability to modify at will. However, that is not really a system, and the cost can be uncertainty, increased need for error checking and, in the worst case, "black boxes" where inexplicable things happen to data. Implementing a structured system will involve some level of rigidity and discipline (either voluntary or imposed) that may well lead to a loss of flexibility but will provide better quality control.

3.6 System Design

The proposed system will:

- continue to rely on Microsoft Excel for data capture and loading to the database.
- be made up of one database (instead of three as now) which will include all FIPS' source data, including the original source (raw) data, any derived variables and any calculated totals or commodities.
- be controlled via a simple user interface (screens) programmed in VB.NET.
- contain all of the business rules needed to derive variables and new commodities and calculate geographic and frequency totals.
- provide standard reports on quality, process management and data.
- provide templates for other reports using SAS Enterprise Guide (EG)[11] that FIPS users can adapt for their specific needs.

- **User Interface**

It was decided that a custom interface would provide a simple portal for the user, and would ensure a structured and controlled environment. This would be a very simple interface with very few screens

[11] SAS Enterprise Guide: a graphical point-and-click software that allows users to create and display data and reports without the need of programming language knowledge.

which will permit the major functions of: loading data from any source, running calculations, running reports, tracking progress and changes to data or calculations (auditing and process control), and extracting data (to various output formats including the standard outputs for CANSIM and SNA). Of the options evaluated, VB.NET was the clear winner.

- **Database**

The two options assessed were SAS and MS SQL/Server. Both are more than capable of handling the task, so the choice will hinge on the choice of software to be used in the calculations. Oracle was the other possible software candidate for the database, but it is simply too powerful and expensive for what is needed for this database.

- **Load function**

The new system will allow the users to easily load any data into the database regardless of the source format (Microsoft Excel, paper, database table). The load function requires four metadata 'keys' to define each data point: the commodity (e.g., wheat), the variable (e.g., production), the time period, and the geographic coverage. Microsoft Excel templates feeding into a VB.NET interface are the most likely option for this. A manual update screen will also be available, with audit trail, for exceptional cases.

- **Business Rule Editor**

The business rules range from simple to very complex. The simplest are deterministic edit rules that deal with variables across one record in one table in the database. As well, there are limit or warning rules, whose purpose is simply to inform subject matter staff to examine the data more closely. The simple rules are easy to implement in any of the potential systems options.

However, the more complex rules deal with donor imputation, i.e., where a value is taken from a different time period or a different geography, and adjusted and used to estimate for a missing value in the current record. In this case, the solution requires the linkage of more than one record and sometimes a loop through a series of records. This is more complex to implement in any software.

The choice rests among structured Microsoft Excel templates, procedural language processing (SAS) and object-oriented processing (VB.NET). Once the software options above have been prototyped and evaluated, the team will confirm the final choice.

- **Analysis and Reports**

No matter what choice is made for user interface or database, both SAS EG and Microsoft Excel will be available for the reporting end of the system. Some standard reports will be implemented using SQL/Server and some as SAS EG templates, but the choice for creating custom reports will be left to agriculture subject-matter specialists. In addition, if they want to enforce some standards for their own process, that will be their decision.

4. Benefits and Data Quality

Once operational, there will be one virtual database system that includes the manipulated final aggregate data from all sources used by FIPS. This database system will allow analysts throughout

the division to access the data using the same tools and techniques, providing a number of benefits in terms of data quality, cost savings and staff moral as follows:

- Update the existing databases to a more sustainable technology (from Visual FoxPro to MS/SQL server).
- Simplify the flow of data between subject matter areas by creating a repository for 'raw' data coming from divisional surveys, other STC divisions and external organizations.
- Offer basic standards of quality by establishing ways of working and offering automated checks, facilitating the flow of data from one work area to another, making verification easier and reducing the risk of errors.
- Provide consistent definitions of important data, so that anyone can easily see the period, geographic coverage, source, data quality and confidence indicators.
- Provide an environment less dependent on individuals and more dependent on teams, thereby less susceptible should a team member not be able to work.
- Lead to the more convenient flow of information and staff. If the information flows are more standardized, it would be easier for staff to move from one work area to another, offering opportunities for their development and cross-training.
- Allow for flexibility and innovation in analysis, by facilitating the export of data to any commercial software that may provide new tools. This certainly includes Microsoft Excel and SAS Enterprise Guide, but may come to include any new desktop software.
- Provide a secure environment that protects the data against inadvertent change but allows updates by authorized users.
- Provide a low-maintenance environment with all of the functions users need to add or change data and reports without the need for systems staff.
- Provide a strong systems commitment to maintaining and protecting the system and data.
- Make documentation and training easily available to users.
- Get business knowledge out of the users and into the system and its documentation.

5. Summary and Conclusions

Canada is fortunate in that it has some of the most detailed information on agricultural income and financial performance in the world. STC has a long history of producing high-quality unbiased information on the performance of the agricultural sector using a multitude of different data sources. Significant improvements have been made over the years on farm income measures to meet evolving user needs and changing industry structure. However, the agriculture industry, along with farm families, is changing rapidly and these changes impact measures of farm income and performance. An effective and efficient system is required to ensure high-quality data are produced, allowing staff increased time for analysis and increased flexibility in adjusting business rules to changes in the industry. The challenge going forward in this major undertaking is to continue to develop and then implement the system as described in the report as efficiently and quickly as possible while continuing to produce the high-quality data and information required by users.

References

Agriculture and Agri-Food Canada (2006). An Overview of the Canadian Agriculture and Agric-Food System, Publication 10013E, May 2006.

Agriculture and Agri-Food Canada and Statistics Canada (2000). Understanding Measurements of Farm Income. Publication No. 2060/B, Catalogue No. 21-525-XPB.

Caldwell, J. and P. Murray (2005). Profitability in Farming: Rates of Return and Comparison to Other Industries. Paper presented at Statistics Canada's Economic Conference, Ottawa, May 9-10, 2005.

Dion, M. (2007). Metadata an Integral Part of Statistics Canada Data Quality Framework. Paper prepared for the Fourth International Conference on Agriculture Statistics, Beijing, China, October 22-24, 2007.

Federal-Provincial-Territorial Working Group on Economic Analysis (2006). Long Term Challenges and Opportunities: Future Competitiveness and Prosperity of the Agriculture and Agri-Food Industry. Progress Report Presented to the FPT ADMs, February 2006.

Murray, P. and D. Culver (2007). Current Farm Income Measures and Tools: Gaps and Issues. Paper prepared for the Agriculture and Agri-Food Canada/Statistics Canada Workshop on Farm Income Measures, Ottawa, March 5 and 6, 2007.

Statistics Canada (2003). Quality Guidelines. Fourth Edition. Catalogue No. 21-539.