

The future of data dissemination

Report on a symposium

The revolution in information technology has radically altered people's requirements for information – not just its content, but its format and speed of delivery. Naturally enough, statistical agencies have been greatly affected by those changed expectations.

Statistics Canada hosted its 12th annual International Symposium on Methodology Issues, November 1 to 3, 1995. Speakers addressed numerous issues related to providing information from many sources of statistical data. The three days offered an opportunity for discussion among the professionals who make, shape, analyze, store, sell or otherwise use statistics.

This report provides brief summaries of selected speakers' remarks about the analysis and dissemination of information: the type of integrated data needed now, the response of statistical agencies to user demands, and the effect of information technology on the distribution of data. All presentations made will be published in the symposium's formal proceedings. (Ordering information follows this report.)

Part 1: The future of statistical data

*Peter Hicks, Head,
OECD project on policy implications of aging societies*

Introducing the keynote address on "the role of statistics in making social policy," Mr. Hicks observed that there is a radical worldwide shift in social policy from an ideological to an empirical basis. This new social policy still needs an

This report was prepared by Susan Crompton, who was with the Labour and Household Surveys Analysis Division.

intellectual/conceptual framework with a statistical database to support it, and Canada is well-placed to play a major role in developing both.

The ideal statistical database should be comprehensive in breadth and depth. Statistics with "breadth" are able to illuminate a wide range of activities such as work, learning, and the family; to analyze intergenerational and lifecycle issues; and to examine the effect of government programs on individual and social well-being. A database with "depth" is strong enough to measure the effectiveness of government programs; to create social indicators to signal emerging problems; and to predict, at the micro level, the best government program interventions for an individual or particular family. The ideal database should include both administrative records and survey data, coupled with the predictive power of microsimulation models.¹

Canada does not yet have consistent and integrated data to track problems over time, although many of the data needed by the new social policy already exist. The difficulty is that the sources cannot be linked since the conceptual framework that would allow their integration has not been formulated. However, Mr. Hicks believes such a framework is emerging. Based on time use – how and where people spend their time – it will offer unprecedented information about behaviour over the lifecycle, investment in human capital, lifelong learning, family life, acquisition of skills, and so on. Integrated time-use databases should soon provide the same solid statistical framework for social policy that Statistics Canada's System of National Accounts now provides for economic policy.

The integrated database, combined with microsimulation models, will allow social programs to be determined by evidence, not by ideology. For example, a program model now being tested by Human Resources Development Canada allows policy makers to assess the effectiveness of different interventions (for example, training or mobility programs) in terms of the real costs and benefits to government and society; at the same time, individuals thinking of participating in a particular program (for example, training or employment programs) can assess the probability of their success, based on the experience of previous participants with similar characteristics.

Eventually, databases such as these might be accessible to individual clients of social services through computer highways or information kiosks, thus shifting program delivery from large government departments to small agencies. Mr. Hicks concludes that using the database in this way could save billions of dollars currently being spent on social and health programs.

*Gordon Priest, Director,
Integration and Development
of Social Statistics
Statistics Canada*

Echoing Mr. Hicks' theme of integrated data, Mr. Priest added that pressures to build broad-based, multi-source information databases – or metadatabases – are growing as informatics technology makes it easier to manage massive amounts of data.

Creating metadatabases requires integrating data by linking sources. However, this has been slow to happen. Statistical agencies have always been "methods driven,"

collecting data using vehicles developed around specific methodologies and serving specialized clientele. Thus, an agency may have evolved as a consortium of relatively independent producers of data, making communication within the organization difficult. As a consequence, agencies themselves are often unaware of the full range of their data holdings, limiting their use to clients; disharmonies may render data from different sources incomparable when subject matter areas develop their own specialized concepts, definitions, classifications and data coding; and the data derived from one source may contradict that obtained from another.

Mr. Priest acknowledged that building and maintaining meta-information databases to accommodate increasingly sophisticated clients will be a challenge for statistical agencies. However, a good metadata base will address the three basic weaknesses described above. It should document the content of the agency's microdata files, its tabular output, and its analytical/descriptive reports; furthermore, this information should be easily accessible using the subject or keyword approach. Resolving every disharmony between data sources may be more difficult to achieve, but the agency should be able to eliminate the worst offences, such as inconsistent classification or coding schemes, by adopting a standardized "template" for survey development. And the analytical outputs developed by the agency should be geared to specific issues or subjects, making use of the full range of relevant information rather than relying on a single source.

*Wouter Keller, Director,
Research and Development
Statistics Netherlands*

Dr. Keller noted that the information technology revolution has

greatly reduced the cost of data processing and storage, thereby changing the way statistical agencies must operate. Statistics Netherlands is already undergoing such a transformation: it is no longer looking exclusively at the internal aspects of data management (editing, processing, organizing), but at the external stages, that is, at data collection and dissemination.

In the Netherlands, medium-sized enterprises receive 30 to 40 statistical questionnaires each year from various government departments. Yet information technology could greatly reduce a firm's response burden, using the same type of computer-assisted data collection techniques that have substantially reduced response burden in household surveys. (This exercise also saves the statistical agency time and money: very little clerical work is necessary to clean data from the household surveys at Statistics Netherlands, but a great deal is needed for the business surveys.)

Statistics Netherlands is proposing to use Electronic Data Integration, or EDI, to collect data directly from each firm's electronic bookkeeping system; in essence, a firm's computer is connected to the agency's and data are downloaded through this link. A pilot study of EDI conducted among a small number of firms in 1994 elicited a generally positive response. However, the real difficulty with EDI linkage is not technological but "linguistic": bookkeeping concepts still have to be translated into statistical concepts, either by the firm or the statistical agency. This "conversion" problem can be solved by an electronic super-questionnaire.

The EDI super-questionnaire would meet the needs of all statistical agencies and government departments sending out questionnaires – on exports, imports, employees and payrolls, for exam-

ple – and reduce response burden to the "touch of a button" that operates the EDI link between the firm's and the agency's computers. But first, statisticians must agree on the definitions and concepts to be used in the questionnaire; then, the electronic bookkeeping used by the enterprise must be converted into the questionnaire's statistical language. Statistics Netherlands is aiming to introduce the electronic super-questionnaire by the year 2000; to this end, it is undertaking a second EDI pilot study of thousands of enterprises in 1996.

The application of EDI technology is not limited to reducing response burden. It can also determine the agency's data output program by integrating and amalgamating statistics in a metadatabase from which information can be downloaded in a variety of formats (CD-ROMs, diskettes, online via the Internet, and other customer-specific products).

Part 2: Customers come first

*Michael Blakemore
University of Durham*

Describing his talk as aggressively pro-customer, Mr. Blakemore called himself a "positioner of information in chaotic times." He reminded his audience that a variety of factors in addition to new technology had pushed statistical agencies to enter the information business. These other factors include government budget deficits, organizational downsizing, privatization, demographics, the switch from income to sales taxes, the globalization of business, and the separation of government policy from its operations (meaning the collector of information is no longer the primary user). All these "push" factors have affected the dissemination of government information and given rise to a number of concerns, especially fear of the

loss of public accountability, stable data series and privacy; worry about the creation of monopolies that can deny access to data for commercial (as opposed to privacy) reasons; and confusion over who – government or business – will be keeping the records.

Based on the experience of several data dissemination organizations in Europe, Mr. Blakemore identified the delivery of specialized services by small niche groups as an area of opportunity. Dissemination practices are changing from the old top-down approach in which the statistical agency develops products, to a bottom-up approach in which the customer defines the product. Furthermore, user groups mutate in confusing and often contradictory fashion, and external influences (for example, reorganization of government departments) can change the ground underfoot in unexpected ways. In these circumstances, he said, distribution systems providing smaller, targeted products can respond more effectively than large centralized systems.

In closing, Mr. Blakemore observed that a large number of customers would still need the services of “gatekeepers” to educate them and monitor their use of the data. He also warned that many people will be denied access to the information highway, often for political, religious or cultural reasons.

*Ulla de Stricker
Stricker Associates*

Introducing herself as an information marketer, Ms. de Stricker said that statistical agencies must recognize that they, too, are in the information business now. Statistical agencies’ customers will demand choice, convenience, and flexibility; they will ask for technological and/or analytical support when they believe it is needed. As they

become more statistically sophisticated, their tolerance for “sacred cows,” such as the painstaking precision that delays the release of data, will disappear. Customers will demand to decide for themselves whether the data quality is good enough for their purposes, and will want to select data from any point on the “information continuum” – from instant but raw, to polished but late. And since each customer must define the content, delivery vehicle, and flexibility of the data required, statistical agencies should allow users to participate in developing products and services.

Ms. de Stricker warned that technology creates new competitors as well as new customer demands. It makes it easier and cheaper to collect and manipulate data, and because the links between buyers and sellers are closer, the market is within easy reach of everyone in the business. In response, statistical agencies must use their information expertise and their data to secure partnerships in the market.

Part 3: Protecting privacy in the era of metadata

*David Brown, Executive Director,
Information, Communications
and Security Policy Division
Treasury Board Secretariat*

Mr. Brown discussed the basic principles of privacy that must be respected in the collection and management of information. People’s fear of losing their “informational privacy” – control over personal information – has grown in recent years, but surveys show that their greatest concern is having this information misunderstood or misinterpreted, leading to decisions that adversely affect them. Data collection for statistical purposes is not generally seen as a threat to privacy.

The federal government is one of the country’s largest collectors, managers and users of data, and as such has been sensitive to privacy concerns for a number of years. It first recognized the importance of informational privacy when it added a section to the *Canadian Human Rights Act*; these provisions were replaced in 1983 by the *Privacy Act*. (Its principles are based on the OECD privacy code, which has been adopted by most Western countries as the basis for their own legislation.) The *Statistics Act*, the *Income Tax Act* and the *Unemployment Insurance Act* also have provisions to govern the handling of personal information in specific institutions. Together, these laws establish the privacy principles that guide the government’s approach to the management of information.

The information privacy principles most applicable to the collection of statistical data require that

- the data collector fully inform the individual why the data are being collected (the purpose, whether participation is voluntary or obligatory, and so on);
- these data not be used for other purposes without the permission of the Privacy Commissioner, who may require that respondents be informed of the new use to which their personal information will be put;
- respondents have the right to access the data about themselves, and to ask that the file be corrected if it contains inaccuracies;
- the institution protect the confidentiality of the data it has collected.

Mr. Brown agreed that the amalgamation of databases to create metadatabases is very tempting, but warned that a balance must be maintained between ensuring the privacy of personal information

and meeting the efficiency goals of information management systems.

*Louise Desramaux, Director,
Data Access and Control Services
Statistics Canada*

Ms. Desramaux introduced her overview of Statistics Canada's privacy policy by reminding the audience of the Agency's two responsibilities: to provide valuable data, and to respect the confidentiality of individual respondents. These two goals can cause conflict but they are in fact linked.

In response to the public's concern about increased accumulation of personal information, and to the legitimate pressures from users to provide more detailed data, Statistics Canada has adopted a legislative-policy framework. Its legal foundation is provided by the *Statistics Act*, the *Access to Information Act* and the *Privacy Act*; the Agency has established internal policies and procedures to ensure compliance. Ms. Desramaux explained that the *Statistics Act* gives Statistics Canada the authority to collect data, obliging respondents to participate in surveys, but also obliging the Agency to protect confidentiality by suppressing any data that could identify a respondent once the data are disseminated. The *Access to Information Act* gives the public the right to access their personal information, but forbids third party access (no one may see the record on anyone else).

The third statute – the *Privacy Act* – has spurred the development of several internal policies that deal specifically with privacy concerns. These include policies on record linkage and informing survey respondents, and the microdata release policy. This last policy was developed in the early 1970s, when the revised *Statistics Act* (1971) allowed the release of anonymous information on individuals. A special committee reviews all pro-

Information technology is making data searches easier at Statistics Canada

*Ross Grenier, Director,
StatsCan Online Project
Statistics Canada*

StatsCan Online is controlled and operated by Statistics Canada in partnership with a private technology company. It provides direct-dial access to detailed data that can be downloaded in a variety of formats: text, tables, and specialized databases. The objectives of StatsCan Online are to improve the accessibility, timeliness and usability of data, including metadata; to increase the breadth and depth of data holdings available to users; to reduce the unit cost of data retrieval to the client; and to improve the cost-to-revenue ratio of data dissemination to the Agency. In the future, Online plans to include mapping capability as well as information from other government departments, and to link to the Internet. Although not officially launched until April 1996, almost 200 clients used Online while it was still in the market testing phase.

*Louis Boucher, Assistant Director,
Dissemination Division
Statistics Canada*

Statistics Canada's Advisory Services receives about 600,000 requests per

year from external users, library staff and Statistics Canada analysts. Using a Windows-based search and retrieval system (known as "Information on our Products and Services" – IPS), employees are able to identify current information about products and services. IPS enables organized and efficient inquiry searches of all "registered" products and services, as well as articles and survey background documentation. Users can query by word, title, subject, author, and date, among other variables. Once the search is completed, IPS compiles comprehensive lists of products and services, allowing staff to download the information, provide lists to clients, or create mini-catalogues.

At the moment, IPS is an internal tool primarily serving staff in Advisory Services and the Statistics Canada library. Plans for the future include making it available to the public via CD-ROM and possibly the Internet. This will hasten the inclusion of more information, including a thesaurus to cover colloquial terms not found in official agency terminology, and linkage to other metadatabases.

posed releases of microdata, and authorizes their release only if two basic conditions are met: they have substantial analytical value, and all reasonable steps have been taken to prevent identification of individual respondents.

Ms. Desramaux acknowledged that people are worried about the power of technology to link data and create massive databases. However, she believes that the legislative-policy framework has been an effective tool in maintaining privacy. Statistics Canada continues to enjoy good response rates, and a 1992 survey on privacy showed that only 14% of respondents were apprehensive about providing personal information to the Agency.

Part 4: Partnerships to serve clients better

*Jan Kestle, President,
Compusearch*

Ms. Kestle described the benefits of a public and private sector partnership from the company's point of view: providing customers with one-stop shopping; enjoying a statistical "license" that enables the integration of data from many different sources; expanding the pool of statistical and analytical expertise; generating increased demand for products; and responding more quickly to the market.

There is a negative side to such a partnership, though, including customer complaints about the

high cost of data stemming from cost-recovery policies at the source statistical agency; controversy about a good collected at public expense being used for profit; competition for clients between the partners; difficulty policing agreements between the partners and their customers; and control of "data leakage" (the piracy of enhanced data).

Ms. Kestle believes that in future customers will demand more value added to the data they purchase; that data will become more accessible though not necessarily better understood by its users; that licensing agreements will be used more extensively to control data piracy; that partnerships will expand in response to the market; that the number of "virtual companies" will increase, making data control and licensing agreements more difficult to police; and that data enhancement costs will continue to grow. All this will necessitate closer relationships and clearer communications among data suppliers if they are to continue enjoying the benefits of their partnerships.

*David Roy, Director,
Marketing Division
Statistics Canada*

Forming partnerships is now a common strategy in the information industry. Because Statistics Canada's strengths are high quality data and analysis, as well as its credibility as a data supplier, the complementary strengths of potential partners would include the ability to solve dissemination problems, to develop sales and distribution networks, and to take risks.

Statistics Canada is currently developing a dissemination framework and network. The next steps

include launching StatsCan Online, overhauling the CANSIM database (which will become the core data warehouse), and converting more publications to electronic formats. To this end, the Agency will adopt partners able to provide online gateways, to co- and re-publish, to develop delivery technologies and to design new products.

*Anne Foster, Senior Vice-President,
Carswell and Thomson
Professional Publishing*

Ms. Foster noted that business partnerships are often compared to arranged marriages of the fifteenth century. In a successful alliance, partners can secure new opportunities; in an unsuccessful alliance, they can lose control of their products, their information or their area of expertise. She identified three factors as critical to the success of partnerships in the information industry: the partners' clarity of intent, or the reason for the partnership; their integrity, or commitment to honesty and plain-dealing; and their interests, which are not the same as intent and must therefore be clearly stated and understood.

Ms. Foster said government and the private sector share only the same intent, which is a healthy economic environment. Thereafter, the similarity ends. The government's integrity rests in not abusing its power, such as using copyright to exclude access to information; the private sector's integrity lies in respecting both the skills and the restrictions faced by the governmental partner. The government's interest lies in disseminating information for the public good, while the private sector's lies in publishing for profit. Having different goals means that the interests must be identified clearly so that partners understand each other.

*Peter Brandon, Partner,
Synovators Ltd.
Editor/Publisher, "Electronic
Information Partnerships"*

According to Mr. Brandon, the world in which the information industry operates is as chaotic as Alice's game of croquet with the Queen of Hearts: the mallets, the wickets and the croquet balls are all live animals, constantly shifting position and changing the rules of the game. Despite this chaos, he believes that new rules for partnerships in the Information Age are emerging.

Partnerships are struck for different purposes – governance, advisory, delivery – and so require different rules. Those established to distribute information should be reinvented because distribution costs are fast approaching zero, and the opportunities to make a living in the field are shrinking.

Partnerships must tap the core strengths of the partners to mutual advantage, and be prepared to rely more on understanding, ethics and codes of conduct than on legal contracts. (Traditional legal concepts may become difficult to enforce as new forms of intellectual property are developed.)

The new partnerships will blur the lines between the company's internal and external operations as firms seek partners to provide goods and services, rather than generating them from within. They will also require more and faster sharing of information, values and expectations. Nor must partners forget that their alliance draws legitimacy from the power of the individual firms.

Mr. Brandon concluded by observing that firms will have to recognize that the human relationship between partners is more important than the technology involved.

Summary

Gordon Brackstone, Assistant Chief Statistician, Informatics and Methodology Statistics Canada

In a brief summary of the three-day symposium, Dr. Brackstone noted that two overriding themes had emerged. The first was technological: the effect of computing power on the collection, processing and management of data, and the effect of telecommunications on the way the data are distributed. The second was financial: the challenges faced

by statistical agencies required to do more with fewer resources, and the extent to which these agencies must cover their costs through data dissemination.

Proceedings of "Symposium 95: From Data to Information Methods and Systems," the 12th Annual International Symposium on Methodology Issues, will be available in summer 1996. Complete coverage of all papers presented is provided. To order, contact Jean-Louis Tambay, at (613) 951-6959; or fax (613) 951-3100.

□

■ Note

1 A microsimulation model is designed to predict the outcome of certain events on individuals with particular characteristics. It does this by creating a synthetic person based on characteristics observed in real persons, and then applying a mathematical model that simulates the probability that specified conditions or events will occur. For example, a simple microsimulation model could be used to predict the likelihood that person A will be a smoker, given the smoking behaviour observed in the general population.