

SEMINAIRE

E+; 3=! C

SEMINAR

STATISTICAL COMMISSION AND
 ECONOMIC COMMISSION FOR EUROPE



Distr.
 GENERAL

CONFERENCE OF EUROPEAN
 STATISTICIANS

CES/SEM.43/19
 8 February 2000

ENGLISH ONLY

Seminar on integrated statistical information
 systems and related matters (ISIS 2000)

(Riga, Latvia, 29-31 May 2000)

Topic III: Innovations in data collection and exchange

**IMPACT OF NEW INFORMATION TECHNOLOGIES ON DATA COLLECTION
 AT STATISTICS SWEDEN**

Invited paper

Submitted by Statistics Sweden¹

I. INTRODUCTION

1. The development of electronically distributed services and new ways of communicating rapidly and easily by means of computers and networks have created new opportunities for the collection of statistical data. This is most obvious at enterprises, where to a large extent, the technology is already in place. For households and individual people, techniques for computer assisted interviewing have been available since the eighties. Individuals and households are still difficult to reach by means of electronic communication but the rapid growth of the Internet has opened up new opportunities. The overall impression of the current status is that techniques such as computer-assisted interviewing are now well established, while the use of web-based services and Electronic Data Interchange (EDI) is only just beginning. Electronic questionnaires can replace ordinary paper questionnaires and in some circumstances data can be collected directly from the information systems of enterprises.

2. The increased use of computers together with government support for an effective communications infrastructure will speed up the process of implementing web-based services in Sweden. The Delivering and Receiving System (DRS) is an example of such an initiative at the governmental level. When fully developed, it will function as an information bus for governmental organisations, enterprises and individuals.

3. To create the necessary conditions for an increased use of web-based services, it is important that respondents feel confident that organisations such as National

1 Prepared by Hans Ireback.

Statistical Institutes (NSIs) treat the data with necessary care. Security issues will therefore become more and more important.

4. This paper describes past experiences with the new technologies at Statistics Sweden, the current status and where we are heading.

II. THE USE OF INFORMATION TECHNOLOGIES IN SWEDEN

5. The last few years have brought a substantial modernisation of technical functions for electronic information services. There has also been an immense increase in the number of users both in enterprises and households who have the technical prerequisites for using these services. Some interesting statistics are presented in the European INFO2000 Member States Study (Statskontoret 1999). The Swedish study on the use of the Internet was carried out in May 1998 and is the most far-reaching survey in the field to date. The survey was conducted by Statistics Sweden at the request of the Swedish Agency for Administrative Development and comprised 6200 interviews with Swedish citizens aged between 18 and 64. The result can be presented by subgroups and regional areas and is, among other things, a valuable source for seeing whether some groups or regions are less well represented than others.

6. The survey shows that 73% of the population have access to a computer at work and/or at home. The number of computers at home has increased sharply. This is partly because most employees have the opportunity to buy a computer through their employer. During 1998 more than half a million computers were bought this way.

7. The survey also shows that over half the adult population in Sweden (51%) has access to the Internet at work and/or at home. This corresponds to about 2.6 million people. The difference between ages is substantial: 80% of those in the 18-19 age group have access to the Internet compared to 30% in the 55-64 age group. The difference between genders is less significant; the figure is 53% for men and 48% for women. The Internet connections are used frequently. More than 34% stated that they download documents and search databases on the Internet at least once per week. Men use the Internet more often than women (40% versus 27%). Self-employed persons are the most frequent users of Internet services such as e-mail, bank services and databases. The pace of developments since May 1998 is hard to estimate. Some smaller surveys indicate that access to computers has increased 7% and to the Internet 15%. The tendency seems to be that the number of new PC-owners is decreasing while more PC-owners are acquiring Internet connections.

8. These figures show that there exists an established base for the increased use of web-based services for individuals and households in Sweden, as well as enterprises.

III. TOOLS FOR WEB-BASED DATA COLLECTION

9. Statistics Sweden evaluates new methods that may improve the process of data collection on an ongoing basis. Improvements can be achieved by using new technologies for data collection and by reducing the reporting burden on respondents. For example, a long-term goal for Statistics Sweden is to offer all enterprises and organisations the possibility of returning information by means of electronic questionnaires. Wherever possible, these questionnaires should also be combined with functions that retrieve data from the respondent's information system. The rapid evolution of the Internet has now reached a phase where the prerequisites are available for applications of this kind. Today the web browser has more or less become the basic user interface. Organisations are starting to use electronic questionnaires, some of them with advanced functions and the potential to communicate with the data collector's database on the server side.

10. An example of this kind of application is Statistics Sweden's web questionnaire for the Consumption of Electricity Survey. This application was introduced in 1999 and developed using traditional HyperText Markup Language (HTML) techniques. It includes validation of input through Java script but also through database communication, and extensive back-end editing via stored procedures in the database. The survey occurs on a monthly basis and input data are checked against previous values. Respondents can obtain several tables with their own data as feedback from the application. Communication with the database is implemented using Microsoft Active Server Pages and the application runs on Microsoft Internet Information Server.

11. However, this traditional approach has certain drawbacks. The lack of standards on the client side makes it difficult to develop applications that work with all common browsers. HTML has been used with such inventiveness that it makes web applications hard to maintain. The use of Java as a standard language for script and components solves some of these problems but requires new skills in the organisation and will not decrease the burden of maintenance. It is therefore natural that organisations search for new tools that will help them to create advanced electronic questionnaires and applications without low-level programming.

12. Statistics Sweden is currently evaluating commercial tools that will facilitate development by hiding as much of the detail as possible from the developer. What these new tools all have in common is that they exploit the possibility of extending the capabilities of Internet browsers by adding components. A component could be a plug-in, Java Applet or a Common Object Model (COM) component. A component could be kept quite small and the time required to download it the first time poses no problems. This will open the way to increased use of advanced and intelligent electronic questionnaires on the Internet.

13. There are several commercial tools designed for the development of electronic questionnaires on the market. One example of COM-based software is FormFlow 99. This has been developed by JetForm, who have offered tools for designing and routing electronic questionnaires for many years. The release of FormFlow 99 marks the company's move away from a proprietary form and data formats to Internet standards.

14. Superform is a Swedish company that has a suite of "fillers" designed for different environments. Their latest product, SuperForm Java Filler, extends Superform's reach to other platforms. The questionnaire is implemented using a Java Applet and its only requirement is that Java Virtual Machine is installed on the client computer.

15. We are also planning to evaluate products that are XML-based. Examples of companies that offer tools of this kind are UWI.COM and the Swedish company Fill-In.

16. Another possible alternative we are investigating is to use standard packages such as Word or Excel to create advanced questionnaires that could be adapted for use over the Internet. Together with the possibility of using traditional Windows applications, this means that we now have a broad range of alternatives for implementing electronic questionnaires.

17. On the server side we have so far used Microsoft Information Server. We have just started a pilot project evaluating Silverstream Application Server. Silverstream is a comprehensive application server that allows corporations to build and deploy complex HTML and Java applications. Silverstream provides support for applications that are deployed to a very large number of users with functions such as load balancing and fail-over. It also provides a multi-tier application development environment with support for COM and Common Object Request Broker Architecture (CORBA) objects. Compared with the Microsoft environment, it has a more integrated tool set.

18. It is impossible to predict the outcome of all these activities, but they will hopefully lead to us having a recommended set of tools by the end of this year. In

order to reach our long-term goal, it is important that this set of tools is flexible and matches our needs, and that it can easily be integrated with our environment.

III.1 Extensible Markup Language (XML)

19. Extensible Markup Language (XML) is best known as a data format for structured document interchange. Like HTML it is a markup language, but XML is much more suitable for describing what kind of information the document contains and how it is organised. Furthermore you can define your own tags and thereby create your own XML implementation. XML is expected to benefit e-commerce by enabling different systems to communicate business transaction information in a known format. One of the problems with Electronic Data Interchange (EDI) standards like EDIFACT (EDI for Administration, Commerce and Transport) is that the definition of the formats is complicated and implementations often require more or less complex installations of special software for converting messages to the internal systems. XML is therefore often mentioned as a replacement for traditional EDI. It is, however, easy to forget that EDI standards like EDIFACT represent a unique knowledge of business processes. Documents and data types represent many years of work by competent people all over the world. EDI has therefore become a common standard, especially for large enterprises, which are unlikely to abandon EDI solutions in the near future. However, XML documents on the Internet offer small and medium-sized enterprises that have not invested in EDI solutions the possibility of electronic data interchange. The XML syntax is easy to use and XML documents can be understood by both computers and humans. Shortcomings like the lack of data-type declarations will probably be solved in the future. XML will also increase the potential for using standard software products. All major web browsers are expected to support XML in the future. Open source frameworks such as Microsoft's BizTalk will house XML schemas (or contracts) for developers using XML. Work on standardisation issues is also being carried out by the United Nations body for Trade and Electronic Business (UN/CEFACT) and the Organisation for the Advancement of Structural Standards (OASIS).

20. Gartner Group estimates that in the year 2003:

- 30% of all EDI will be XML-based
- 30% of all EDI will be handled via bridges between XML and traditional EDI (EDIFACT/ANSI x.12)
- 40% of all EDI will be traditional EDI.

21. XML can also be used to define validation and navigation rules for electronic questionnaires. One example of such an initiative is eXtensible Forms Description XFDL. A proposal that will address this issue (IQML) has also been accepted within the EU's 5th Framework Programme.

22. The combination of XML and Java is very likely to become a common base for new technical solutions and services. XML offers technique-independent data exchange and Java offers technique-independent code execution. Together they can make it possible to build solutions involving intelligent electronic questionnaires that do not need frequent and time-consuming communications with the web server.

IV. SECURE ELECTRONIC COMMUNICATION

IV.1 Major considerations

23. The three major considerations for effective IT security are confidentiality, authentication and integrity. These issues are discussed further in the following sections.

IV.1.1 Confidentiality

24. Confidentiality means that information is unavailable to those who are unauthorised to access it. The importance of confidentiality can vary in different surveys. For example, in the case of Structural Business Statistics the concept of confidentiality is extremely important. If an NSI allowed a breach of confidentiality and enterprise information was stolen, the NSI would lose credibility and the enterprise could be harmed.

IV.1.2 Authentication

25. The most common implementation of authentication is the use of passwords and the most common form of security breach is the compromising of these passwords. Using strong authentication and electronic identity cards or single-use password devices are some of the steps that can be used to prevent unauthorised access to sensitive information. Strong authentication is performed by a challenge/response mechanism, which has the property that verification does not require knowledge of a password or any other secret information.

IV.1.3 Integrity

26. Integrity ensures that information cannot be modified in unexpected ways. A loss of integrity could result from human error, intentional tampering, or other external events. The consequences of using inaccurate information can be serious. This issue also involves measures to prevent repudiation, i.e., to prevent senders from claiming that they did not actually send the message. This technique, which involves encrypting a piece of data with a private key, is called a digital signature.

27. A security strategy should be based on all three of these considerations. Depending on the individual needs of the survey, various levels of emphasis should be placed on each aspect.

IV.2 Public Key Infrastructure (PKI)

28. The PKI is the most common security technology for web-based services on the Internet. There is support for the PKI in standard browsers such as Navigator and Internet Explorer through the SSL protocol (Secure Sockets Layer). The PKI provides mechanisms based on public key cryptography to support security services such as authentication, integrity and confidentiality. To understand how the PKI functions, a basic knowledge of the mechanisms involved is required.

IV.2.1 Encryption

29. Encryption is the process of applying an algorithm to a message. The algorithm scrambles the data and makes it very difficult and time consuming to deduce the original if only the encoded data is available. Inputs to the algorithm typically involve additional secret data called **keys**, which prevent the message from being decoded even if the algorithm is publicly known.

30. The strength of the encryption is dependent on the nature of the mathematical algorithm and the size of the keys involved. Under United States regulations, the length of the key that can be used in exported software is limited. Unfortunately the 40-bit encryption limit that has been in force until recently has proved to provide little security from attack. Each extra bit in the key doubles the time needed for an attack and most experts now claim that 128-bit keys are required to ensure reasonable confidence for vital areas such as electronic commerce. Many non-US companies have now developed add-on cryptographic products, using 128-bit or longer keys.

31. On 16 September 1999, the White House announced a new encryption policy allowing for the export of software of any key length to any country except for, what the United States consider as, the seven state supporters of terrorism. This new policy was expected to be published on 15 December 1999 but has been postponed until 14 January 2000. Once it is published, software developers expect to start making products and product updates that contain strong encryption and will be available to customers world-wide.

IV.2.1.1 Symmetric cryptography

32. In symmetric cryptography the encryption algorithm requires the same secret key to be used for both encryption and decryption. The advantage of these algorithms is that they are fast and efficient. However, the problem is that of key exchange. There must be a mechanism for safely ensuring that both parties, the sender and the receiver, have the secret key.

IV.2.1.2 Public key cryptography (Asymmetric cryptography)

33. One solution to the problem of key security, used in the PKI, is asymmetric cryptography. This process uses two keys that are mathematically related. One key is called the private key and is never revealed, and the other is called the public key and is freely given out to all potential correspondents. A sender uses the receiver's public key to encrypt the message. Only the receiver has the related private key to decrypt the message.

IV.2.2 Digital signatures

34. Digital signatures involve swapping the roles of private and public keys. If a sender encrypts a message using his private key, anyone can decrypt the message using the sender's public key. A successful decryption implies that the sender, who is the only person in possession of the private key, must have sent the message. This also prevents repudiation, that is, the sender cannot claim that he did not actually send the message. A piece of data encrypted with a private key is called a digital signature. Common practice is to use a message digest as the item of data to be encrypted. A message digest is a digital fingerprint of a message, derived by applying a mathematical algorithm to a variable-length message. There are a number of suitable algorithms, called hash functions. A message digest can be used to guarantee that no one has tampered with a message during its transit over a network. Any amendment to the message will mean that the message and digest will not correlate.

IV.2.3 Digital Certificates and Certificate Authorities

35. A digital certificate is an item of information that binds details about an individual or organisation to their public key. To make the public key certificates generally available, they should be published in a network directory. The most widely accepted format for digital certificates is the X.509 standard, which is applicable to both clients and servers. With access to someone's certificate and the public key, it is possible to carry out secure communication as described above.

36. An organisation that creates and signs certificates is called a Certification Authority (CA). A CA is a commonly known trusted third party, responsible for verifying both the contents and ownership of a certificate. The term Certification Service Provider (CSP) can also be used for this kind of organisation. Digital certificates include the CA's digital signature, i.e. information encrypted with the CA's private key. This means that no one can create a false certificate. The public keys of trusted CAs are stored for use by applications like web browsers. If for some reason a certificate has to be revoked, it will be added to a Certificate Revocation List (CRL). The CA must maintain a CRL corresponding to the certificates issued. One CA's certificate can be used to verify other CAs' certificates. This is called cross

certification of certificates and makes it possible to use a top node to verify certificates from different CAs. This technique could be used for reducing the problems involved in managing keys from different CAs.

IV.2.4 The electronic identity card

37. A very secure holder of information and private keys is the electronic identity card (EID card). The EID card is based on a smart card, i.e. a plastic card equipped with a microprocessor chip. The microprocessor has a tamper-resistant memory containing private cryptographic keys and digital certificates related to these keys. This makes it possible to use strong authentication, decryption and digital signatures. It employs digital technology to secure digital information in an analogous way to the securing of ordinary paper documents by envelopes, seals and hand-written signatures.

38. The private key stored on the card is only used in the microprocessor's internal operations and is never exposed outside the card. It is not possible to copy a particular card by analysing the inputs and outputs. Digital signatures and authentication based on EID cards can therefore be regarded as very safe.

39. Access to the EID card is protected by a short secret code called a PIN (Personal Identification Number). There are also products on the market that use biometrics methods, e.g. fingerprint technology, as an alternative to the PIN code. In the future it will also be possible to use the EID card in combination with mobile phones. This will open up interesting new possibilities if combined with web services based on Wireless Application Protocol (WAP).

40. Standards for cards and card readers are promoted by the PC/SC working group and Open Card Framework.

V. CURRENT SITUATION IN SWEDEN

V.1 CAs and EID cards

41. The Swedish Agency for Administrative Development helps to develop Swedish administrative policy and is promoting work to ensure that the electronic infrastructure in the public sector is open and secure. One of the agency's tasks is to draw up framework agreements for IT and telecommunications use in public administration. One example of such a framework is the agreement with the Swedish Postal Authority and Telia (a telecom operator) on secure communications services for public services. The Swedish Postal Authority and Telia are both acting as trusted third parties. They are CAs and issuers of EID cards.

42. The growing interest in the possibility of exchanging information securely over the Internet has so far led to most widespread implementation in the area of electronic banking. One of the major Swedish banks, Nordbanken, started using EID cards in December 1997. During 1998 the bank provided over 39 000 customers with EID cards.

43. One of the problems currently under discussion is the management of certificates from different CAs (cross certification). For instance, certificates from the banks are not publicly available as they are handled internally by the banks. One solution discussed is to create a top node that identifies the organisations that act as CAs. Several governmental authorities have been suggested as top node. However, it is not clear whether this top node should be the responsibility of a governmental authority and discussion still continues. An alternative could be an accreditation organisation that is used in other contexts, e.g. for ISO certification.

44. It is also important to coordinate this work with expected developments in the EU Member States. A new EC directive sets a framework for digital signatures. This directive deals with the use of certificates issued in different Member States. The new EC directive is expected to become law in Sweden in 2001. Sweden is also active in international standardisation work through the non-profit organisation Secured Electronic Information in Society (SEIS).

V.2 Internet applications at Statistics Sweden

45. In the web applications that Statistics Sweden has developed so far we have used Secure Sockets Layer (SSL) security. The respondents at organisations or enterprises use passwords as authentication. SSL is activated through a certificate that is installed on our web server. On the client side SSL is supported in both Netscape Navigator and Microsoft Internet Explorer. However, our current server certificate supports only 40-bit encryption. Since 1997 some banks have obtained permission to use stronger encryption using a 128-bit key-length. As described earlier in this paper, the US Government expects to allow the export of keys of any length in the near future.

46. In the meantime, we are working on a test implementation of a PKI solution with non-US cryptographic software. The survey being used is the Structural Business Statistics, which offers a subset of the enterprises (1000) an Excel questionnaire instead of the paper questionnaire. This approach will have the following properties:

- It will be possible to download the questionnaires from our web server.
- Different levels of security will be offered. For some surveys it is sufficient if only the data (the answers) are encrypted for sending to our server. If a higher security level is necessary the whole questionnaire can be encrypted and provided with a digital signature to achieve non-repudiation.
- It will be possible to use certificates and EID cards.
- The server software will be able to communicate different messages, e.g. editing errors, from our database to the client questionnaire.

47. This pilot will use Excel, but the software required for adapting the Excel questionnaire can be used with other types of tools since its functionality is exposed through an Application Programming Interface (API).

V.3 Reducing the respondent's burden

48. A promising approach for reducing the burden of administrative reporting on enterprises is to use information already available in the enterprises' information systems. An EU project called TELER (TELEmatics for Enterprise Reporting) has proved that this concept is viable. There are, however, a number of problems that remain to be solved. One of them is the need for harmonised metadata at the data collector and the enterprises. In Sweden, we have the option of using the BAS system, which is a widely used system for standard accounts. This makes it possible to map statistical variables onto accounts in enterprises' book-keeping systems. In the TELER trials, Statistics Sweden used both the prototype EDISENT, developed within the project, and our own software, called IFK. The BAS standard makes it possible to define the mapping of statistical variables onto accounts before the questionnaire is distributed. The actual extraction of data from the book-keeping system to fill in the questionnaire is done via a file at the enterprise. The respondent can verify the information in the questionnaire and add entries for variables where no mapping is defined.

49. We are now developing the mapping module further to make it work with different types of questionnaires. We have recently started work on adapting the Excel questionnaire used in the Structural Business Statistics Survey. Our plan is to conduct trials with this adapted questionnaire later this year. The enterprises

needed for these trials will be selected from the 1000 enterprises that already are offered the Excel questionnaire.

V.4 Delivering and Receiving System (DRS)

50. The broad acceptance of the Internet makes it possible for public organisations to offer increased services to citizens and enterprises. A necessary prerequisite is that the information exchanged will be standardised and secured. The DRS was originally an initiative from a working group called Top Managers Forum, led by the Minister of Finance. As the work has continued, the Swedish Agency for Administrative Development, the National Social Insurance Board and the National Tax Board have assumed leading roles. Statistics Sweden is a member of the DRS Forum and the DRS Working Group. DRS is expected to become an important part of the Swedish information infrastructure in the next few years. The costs for managing the exchange of information between public authorities can be reduced drastically when the process can be standardised and carried out via the Internet. One of the main functions in the DRS concept is the standardised label that is attached to every document or message sent. This label is created as an XML document and contains information about the sender, receiver and type of document. DRS does not require standardisation of the documents or data files that are exchanged. It is therefore technically possible to exchange all types of documents. However, this means that the communicating parties still need to agree on the data formats used in exchanging information.

51. The potential exists for a number of new electronic services for citizens, enterprises and public authorities. An important element of DRS is the integration with security services based on agreements involving CA services and EID cards. The first version of DRS is now available and supports secure communication between governmental organisations. Future versions will also extend to enterprises and citizens. For instance, in the future (in 3-4 years' time) the Tax Board expects that a majority of Swedish taxpayers will use the Internet to submit their tax returns. This implies that the concept of EID cards will achieve broad acceptance in the society.

SHS is also expected to help the public authorities in their efforts to collect specific information from a respondent only once. By defining agreements in the DRS system, information collected from a respondent can be distributed to several different authorities.

52. It remains to be seen if the expectations of governmental initiatives like DRS, CA top node and EID cards will achieve widespread acceptance. An NSI like Statistics Sweden is dependent on major authorities such as the Tax Board and the National Social Insurance Board acting as the driving forces in development. These authorities provide the services that will motivate enterprises, organisations and individual persons to acquire the necessary components, such as certificates and EID cards. However, even without this infrastructure it is possible to develop and use Internet data collection applications on the basis of existing environments and tools. For many surveys a standard SSL solution with password authentication will be sufficient. The new encryption policies announced will also create the potential for increased security for these types of applications.

53. The rapid evolution of tools for developing Internet applications together with the broad acceptance of the Internet yields exceptional prospects for the widespread use of web-based services. This will mean cost savings and other benefits for all actors involved.

References

Elektroniska informationstjänster och mjuk infrastruktur i Sverige. Statskontoret 1999:40

Struktur för hantering av certifikat och kryptonycklar. Statskontoret 347/98-5.

Användning av Elektroniska ID-kort (EID). SEIS - Secured Electronic Information in Society, 1998.

TELER. Final report AD1016 TELER. Deliverable 08.

XML på den amerikanska scenen. Sveriges tekniska attachéer 1999.

ebXML. (UN/CEFACT, OASIS). <http://www.ebxml.org>

BizTalk. <http://www.biztalk.org>

W3C. <http://www.w3c.org>