STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 14
English only

CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE/EUROSTAT Work Session on Electronic Data Reporting
(Geneva, Switzerland, 13-15 February 2002)

Topic (ii): Security, confidentiality and privacy issues

EXPERIENCES WITH SECURITY IN STATISTICS FINLAND

Submitted by Statistics Finland[1]

Contributed paper

ABSTRACT

This paper describes the experiences of Statistics Finland concerning security in electronic data collection. It gives an overview of the structure of the data collection in Statistics Finland, informs on current technologies and their lifecycles as well as gives some views about the role of security in electronic data collection.

To support these views, the paper also presents some of the basic security requirements for internet-based data collection through a real-life case-study on how Statistics Finland has addressed security issues in internet-based data collection of the building cost index system. Then the paper touches on the issue of using PKI (public key infrastructure) as a potential solution for some of the problems in internet-based data collection and evaluates the applicability and current status of this technology in Finland at the moment.

I. INTRODUCTION

1. Transparency, and efficiency and easy-to-use application are the driving forces of the vastly increased usage of internet based services. No longer is it necessary to visit a supermarket in order to buy goods, and even transactions with official authorities can be performed 24/7 via Internet. However, there are risks in the networked world, many of which concern privacy and integrity. Reactions to these risks vary, but we are coming to a stage where one often must decide on the optimal position between usability and security - and this choice should be a conscious one.

2. Statistical offices collect data in electronic form in order to increase the quality of statistics, to lower the burden of the respondents and to speed up the statistical processes. Although it might be tempting to build up a data collection system from off-the-shelf packages and start it up as soon as possible, it is very dangerous to carelessly exploit these almost infinitive possibilities without a regard to the security issues – the very cornerstone on which the trust that enables direct data collection is built.

[1] Prepared by Sven Björkqvist (Statistics Finland, Data Administration), Eero Koljonen (Statistics Finland, Information Security), and Toni Räikkönen (Statistics Finland, EDP-Methods).

II. THE STRUCTURE OF STATISTICS FINLAND'S DATA COLLECTION

3. The Finnish government has lately been emphasising the importance of electronic information interchange between individuals, enterprises and authorities. Among the authorities is, of course, Statistics Finland, which has managed to convert a large part of its data collection to electronic data interchange.

4. Indirect data collection plays an important role in Statistics Finland. We receive c.a. 94 percent of our raw data[2] from other authorities such as Tax administration, social insurance institution, vehicle register and so forth. The direct data collection covers only about 6 per cent of our data collection, consisting of traditional paper questionnaires (2%), machine readable data/Primary EDI (2%) and interviews (2%).

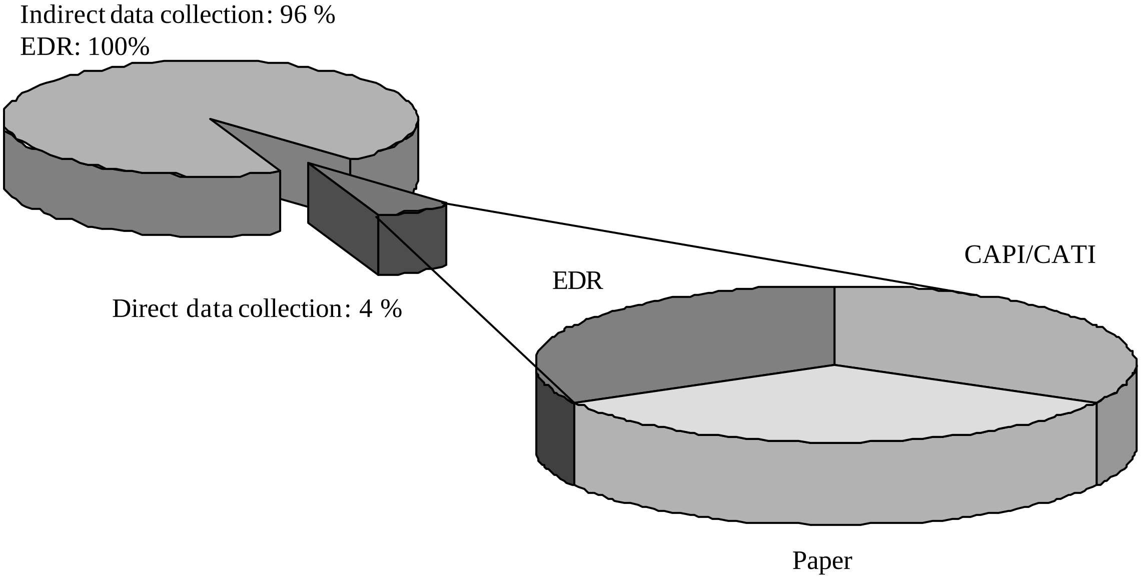


Figure 1. Data collection in Statistics Finland by type and media used

5. Although most of our primary EDI transmissions use other mediums than the Internet, we are focusing on the use of the Internet because Finland has the potential to do so: we have one of the worlds' greatest rate of Internet connected computers per capita. In practice this means that almost every enterprise, school and many households have the possibility to connect to the Internet, thus meeting the first requirement to use it as a medium of reporting to authorities.

6. The percentage of primary EDI in our data collection is only about 2%, but of our whole data collection EDI covers nearly 98 per cent, as the data from other authorities and interviewers is sent to us in machine readable form. In practice this means that only 2 per cent of our raw data collection is based on traditional paper questionnaires (Vallaskangas 1998).

III. THE TYVI-MODEL

7. In Finland the Ministry of Finance, Statistics Finland and other authorities have agreed upon using a special infrastructure for gathering data from the enterprises. This architecture is called the TYVI model (Finnish for data flows from enterprises to authorities). TYVI model consists of several instances, but the main model is divided into three parts: the enterprises, The TYVI operators and the authorities. The model itself describes how data flows from enterprises via operators to authorities.

[2] Raw data consists of all the data sent whether it is used or not.

8. The need for such a model rose from the fact that while the expenses involved with paper questionnaires were diminishing, due to the use of electronic questionnaires, there was also a new field of expenses: the data communication mediums and especially the costs related to the development and maintenance of the several bilateral interfaces needed for gathering the data (see Figure 2). Another reason for the creation of TYVI model was that once data is in machine readable form it is very cost efficient to distribute it amongst the authorities - thus diminishing the burden of the data providers.

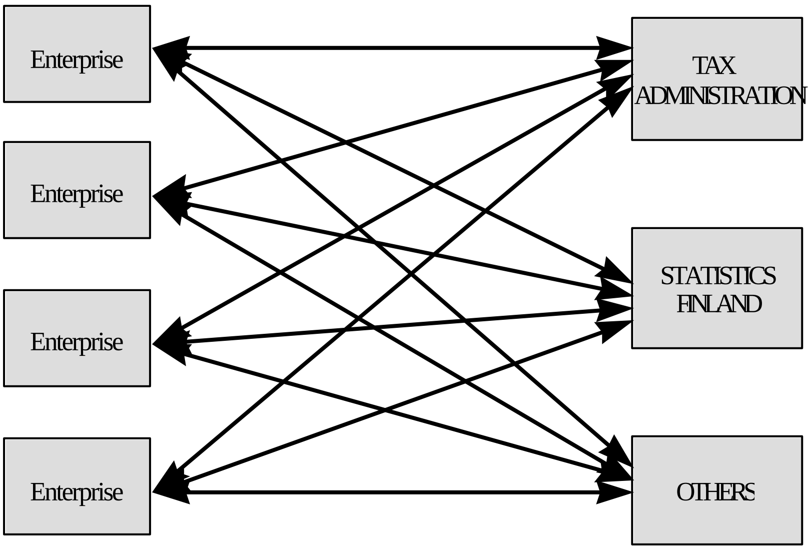


Figure 2. Before TYVI: Model based on bilateral data communications interfaces

9. The essential advantage of the TYVI model is that enterprises and authorities do not need to change their telecommunication interfaces to match each others', as the operators take care of the conversions needed. The authorities and the enterprises do not need to use the same operator, since TYVI operators are allowed to exchange information between themselves in order to provide it to their customers (authorities). Another significant advantage is that by using so-called combi-questionnaires (or a set of questionnaires) authorities can lessen the burden of the data providers as no data is gathered twice. This can be achieved through cooperation of the authorities using the services of the TYVI operators in the following manner:

?? the authorities agree on the information they require;

?? the authorities and enterprises agree that the operator can provide the data needed by each authority, extracting the data from a single (not authority specific) response from a data provider;

?? the authorities (or the operator) put together an electronic questionnaire or a set of questionnaires requesting the data needed;

?? the enterprises fill in these questionnaires and send the data (for instance via the Internet) to the TYVI operators of their choice;

?? the operators gather the data and send to each authority the appropriate data.

10. The primary advantages of the TYVI model to enterprises are:

?? in an ideal situation the same data is reported only once - not separately to all authorities who require it;

?? the interface to all the authorities is the TYVI operator of the enterprise's choice (operators exchange information between themselves so that if an authority and an enterprise each use a different operator, the data is still transmitted);

?? enterprises do not need to adapt their systems to match those of the authorities;

?? enterprises do not need any special knowledge in data communications as the TYVI operator deals with most of these as part of the services to enterprises;
?? the enterprises may freely make contracts with any TYVI operator and they may together decide how the data is transferred to the operator (excluding specific demands deriving from the use of certain software modules).

11. The primary advantages to the authorities are:

?? authorities can obtain the directly collected data in electronic form without expertise in electronic data collection or telecommunications;
?? the TYVI operators take care of the password management and related issues;
?? the authorities do not need to adapt their data communication systems to match those of the enterprises;
?? the TYVI operators provide support to enterprises in cases of telecommunication problems;
?? the data can be converted to a form which best suits the needs of each authority.

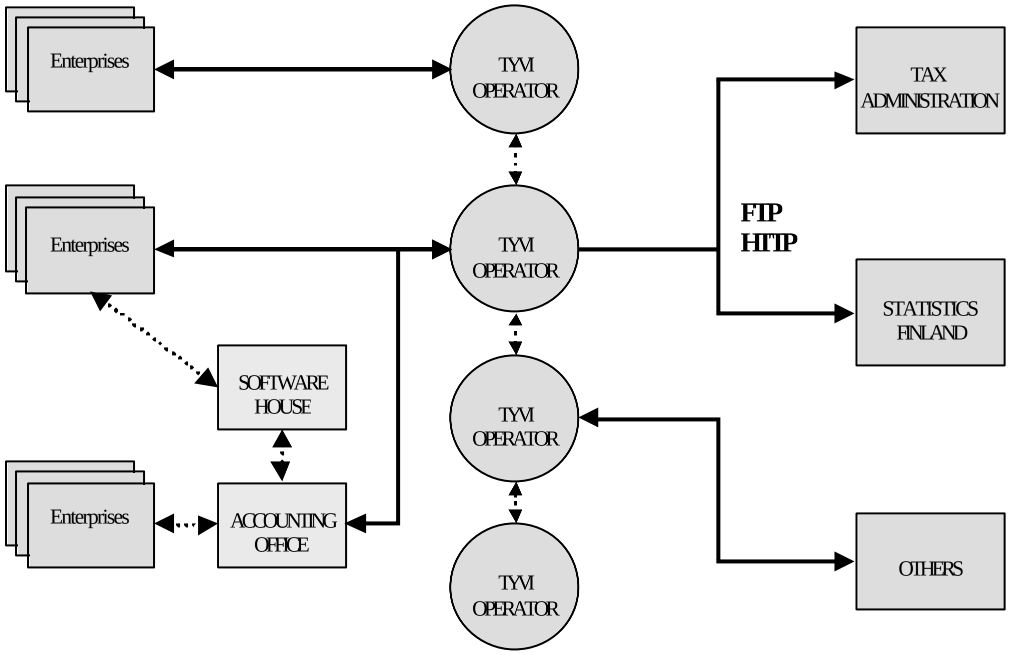


Figure 3. The TYVI -model

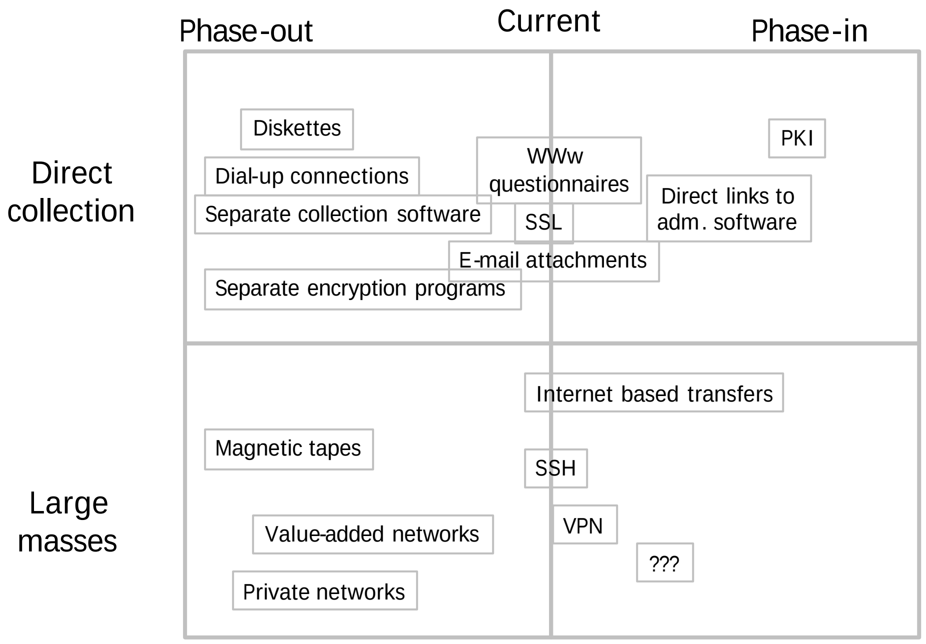
IV. GENERAL EXPERIENCES WITH DATA COLLECTION USING THE INTERNET

12. Statistics Finland has actively tried to increase the use of Internet as a medium for data collection. Maybe the best example of this is the questionnaire for compulsory schools. Compulsory schools can answer this obligatory questionnaire only by using Internet and a collection of WWW-based forms designed for this questionnaire. This means that over four million data elements, consisting of the responses of 4,300 compulsory schools in 436 municipalities, are transferred through the Internet to Statistics Finland using the TYVI infrastructure. This method of reporting is seen as extremely positive by the data providers (compulsory schools) since it streamlines the process of data provision through the use of new technologies. Because of the success of this questionnaire and other positive experiences with questionnaires implemented in WWW-forms, Statistics Finland will probably increase the number of this kind of questionnaires.

13. Statistics Finland has also had other experiences with EDI using the Internet. Many electronic questionnaires (produced with a variety of tools) use the Internet as a medium for transporting the responses from data providers to Statistics Finland. The most important questionnaires of this kind are described next.

14. One of the largest questionnaires with possibility of reporting in electronic form using the Internet is the MS Excel-based annual questionnaire for banks and credit institutions. This questionnaire contains information that is needed by three different authorities: Statistics Finland, Financial Supervision Agency and the Bank of Finland. Therefore, the questionnaire is built so that, once completed, it selects the information needed by each authority then saves and encrypts this information in an authority-specific file. The encrypted reply files are sent to authorities as an attachment of an email message using standard Internet mailing facilities (SMTP). There is also a possibility to copy the response file to a diskette and send it by ordinary mail, but it is intended only as an alternate medium for those respondents that do not have the possibility to use the Internet. In the near future the use of electronic mail and diskettes will probably be replaced by the use of HTTP or other protocols that make it possible to fully exploit the advantages of the TYVI model and the services of the TYVI operators, such as immediate validity checks (during the sending session) on the data sent and real-time feedback to the sender.

V. TECHNOLOGY DEVELOPMENTS



VI. EXPERIENCES IN INTERNET SECURITY

15. This section is structured so that it first examines security requirements in general for Statistics Finland and then concentrates on the specific points which need to be taken into account when operating in a hostile Internet environment. A case study of our BCI-system is presented in the latter part.

VI.1 General security requirements

16. When considering security issues that may arise when collecting data in traditional ways, two points come to mind. First, for additional processing it should be verified that the collected data is reliable. Also, most of the time the collected data is strictly confidential, so it should be taken for granted that raw data is kept in a safe place and handled only by those persons with access authorisation.

VI.2 Security requirements when operating in Internet

17. When operating in the Internet environment, the issues mentioned in the previous chapter also apply. In addition, some new issues also arise:

?? the data can be modified before it reaches the destination (reliability);
?? the data can be grabbed by outside (confidentiality);
?? the respondent can be faked.

18. Threats focused on data can be solved with secured connection protocols and encryption. To verify that the respondent really is who he claims to be requires thorough authentication.

VI.3 Case: building cost index–system

VI.3.1 General

19. Building Cost Index describes new construction in the building trade and is calculated as the weighted average for blocks of flats, attached houses, office and commercial buildings, and for industrial production buildings and warehouses. The Building Cost Index is calculated according to Laspeyres' price index. In addition to the weight structure of the base year, index calculation requires monthly price monitoring of selected commodities. The price changes of commodities are monitored in the compliance with the principles of the so-called pure price index by eliminating the portion produced by quality changes from the price changes detected.

20. The price data are collected from the respondents monthly and are mainly collected from wholesale traders, directly from manufacturers and retail traders. Traditionally, the data is collected in paper form. With this background we wanted to develop the system that would make data collection more comfortable for respondents than the paper form.

VI.3.2 Planning and implementation phase

21. It was very clear at an early stage that data that was passed between Statistics Finland and its respondents was classified. So security was the main task during the creation of this system. Firstly, the connection between respondents and Statistics Finland had to be secured. In practice this was achieved using SSL-protocol (Secure Sockets Layer) during the entire session. The authentication itself was made form-based. This, somewhat but not entirely, prevents so called brute-force attacks that can be made with NT-based authentication. The authentication itself was made with data encrypted with RC4. As the algorithm itself isn't very strong, in this case it's sufficient enough when tunnelled with SSL.

22. Logging was also implemented in the collection system itself. Once the respondent logs into the system, his IP-address is registered. If during the session the IP-address changes from the original one, the respondent will be kicked out with a message implying that there might be a possible security breach.

VI.3.3 Deployment phase

23. In the deployment phase the actual web-server was first secured. When operating in Windows NT platform this included applying the latest service packs, security fixes for the Internet Information Server (IIS) and also hot fixes for the NT core itself.

24. After that the file system protection took place. In general one should always keep the actual application data separate from the operating system data, for example by using different physical partitions for each of these. The actual file system protections were done mostly by following the

instructions from Microsoft itself. As a rule of thumb, it is always a good practice to avoid default installation locations.

25. After that the windows registry was protected. Any so called “dangerous” entries were removed and some others modified. As a last stage all the unnecessary windows’ services were removed as were the unused protocols.

26. After the server was secured it was connected to the DMZ (demilitarized zone) and the rules for accessing the server were included in the firewall configurations.

27. When the deployment phase was finished a third party security company was hired to audit the system. Some suggestions were made and some extra protection applied.

VI.3.4 Experiences and improvement suggestions

28. Experiences with the pilot system have been very promising. Some general questions still arise that are worth considering in the electronic data collection system. One is key management especially in production, distributing and inventorying keys. Statistics Finland produces the keys centrally in order to confirm the robustness of the keys. The keys are randomly produced. The length of code is 8 bits.

29. At this stage the codes are individually based but they can be modified to session based too. The session based keys must be critically considered in statistical data collection because they most likely increase the burden of response and affect response negatively. If the key is lost, the old one must be revoked and a new one produced and, if this happens frequently, it leads to difficulties in administration.

30. The key distribution problem is also noticeable if the data collection extends to thousands of firms. The experiments with electronic data collection in Statistics Finland have been limited to some hundreds of companies and have been easy to maintain.

31. A gradually vanishing problem is the old versions of browsers. Only the latest versions support robust encryption. Statistics Finland is able to support strong security of 128 bites. We generally, however, have to be satisfied with the weak security offered by 40 bites, due to the old versions of browsers of the firms.

32. It does seem, though, that the respondents are far more willing to give their information electronically than in the traditional paper form. It’s also far more comfortable for our statisticians to handle the data when they don’t have to try to decipher what is written on paper. The data will be validated instantly when the respondent sends the information and if something doesn’t pass the validation system, the respondent will be immediately informed and given a chance to correct the information or explain why it differs so much from previous data.

33. The feedback from the respondents has also been very encouraging and some noteworthy suggestions to improve the system were received.

VII. CAN PKI SERVE AS THE ULTIMATE SOLUTION?

34. Many of the problems mentioned above are likely to be solved when we can exploit the public key method, where a third, trusted party (CA) provides the key codes and confirms their security. The public key method provides a pair of keys, the public key for encrypting and the private key for decrypting. The keys are inverse and furthermore asymmetric. All that you encrypt with the public key can be decrypted with the private key and you cannot disclose the one though the other is generally known. That is why the public key can be freely distributed.

35. A public key method could be applied to web-enquires. Then there is no need to distribute identification codes and passwords by post to respondents. You can freely use a public key of Statistics Finland to encrypt the data and send it safely through the network. You can authenticate yourself by signing your message digitally with your private key. Furthermore, passwords and the identification codes need not be saved on the server because the CA is responsible for the safe storage and distribution of the keys, etc.

36. In Finland there have been great expectations of the implementation of the PKI, especially in the government sector, but they are too early. The public key method has not yet been implemented, and the PKI architecture is not ready or standardised.

37. The implementation of PKI causes expenditures to respondents in the form of equipment as card readers and software. The basic demand in data collection is that the users can exploit the encrypting method with ease and simplicity. Furthermore, the PKI should be transparent and applicable to other applications as e-mail, the Internet, use of electronic forms and encrypting files and folders. That is why the PKI must be transparent and standardised. There is no need for users to understand how the PKI operates the keys and certificates in encrypting and digital signing.

38. There remain several uncertainties with the PKI and it will take some years until it is a noteworthy option in electronic data collection. The SSL typed data security can be assessed adequately in data collection until the problems of the PKI method have been solved.