STATISTICAL COMMISSION and   **COMMISSION OF THE**
ECONOMIC COMMISSION FOR EUROPE   **EUROPEAN COMMUNITIES**

CONFERENCE OF EUROPEAN STATISTICIANS   **EUROSTAT**

Joint UNECE/Eurostat Seminar on Integrated Statistical
Information Systems and Related Matters (ISIS 2002)
(17-19 April 2002, Geneva, Switzerland)

Topic II: Secure communications and data confidentiality

ELECTRONIC DATA REPORTING FROM MUNICIPALITIES AND BUSINESSES.

SOME EXPERIENCES FROM TWO PROJECTS IN NORWAY

Contributed paper

Submitted by Statistics Norway [1]

I.   INTRODUCTION

1.   In recent years more and more municipalities and businesses have obtained access to the Internet. This is one of the reasons for Statistics Norway (SN) deciding to give them the opportunity to send in their forms electronically instead of the more standard paper based forms. Electronic data reporting will hopefully save SN both money and work hours compared to today's paper based data reporting.

2.   Statistics Norway (SN) has two major projects in progress involving electronic data reporting from external agencies and companies. These two projects are called KOSTRA and IDUN. This paper will give a short overview of both these projects before presenting more in depth some security and confidentiality issues related to the projects.

3.   Statistics Norway has an ongoing collaboration with "The Directorate of Taxes" and "The Brønnøysund Register Centre[2]", in order to ensure that all agencies can benefit from each other's experie nce. The IDUN project in particular has been developed in close contact with the other two agencies in order to coordinate online user-interfaces, and demands for equipment and technology. This collaboration has had a future goal to develop a common web-solution for all the information the three agencies receive over the

---

[1] Prepared by Magne Hopland.
[2] "The Brønnøysund Register Centre" is the Norwegian registration department, and consists of among others "The Register of Business Enterprises", "The Register of Company Accounts", see also: http://www.brreg.no/english

year. The target has thus been to ensure that no business company or municipality needs to deliver the same information more than once.

A.   IDUN[3]

4.   IDUN started in February 2000 and has as its main goal to develop a scaleable, general solution for web-based data collection from business companies. The purpose is to enhance the quality and speed of the exchange of information with business companies in order to reduce businesses workload, both real and emotional. By developing an electronic report-chain system, Statistics Norway wants to give each client:

- Feed-back to the clients on their own statistical information
- Simple statistical analysis of the clients data
- Links to relevant statistics or market information
- Basic information about the enterprise or business in the Enterprise and Business Register, and an opportunity to update information located there

B.   KOSTRA[4]

5.   The KOSTRA project started in 1994 and has as its objective to contribute to a coordinated and improved reporting chain between Norwegian municipalities and counties and the Norwegian state. To achieve this objective, Statistics Norway has been implementing a solution to replace the existing paper based report forms with electronic forms. By doing so Statistics Norway mainly hopes to achieve two benefits:

- Enhance the quality and consistency of data and statistics on the use of resources in local government administrations in Norway, and to improve comparability between information from different local administrations.
- Collect, compile and disseminate statistical information in this field with fewer resources in terms of money and manpower.

The KOSTRA project is currently in a Pilot phase, but is from the summer 2002 going to be a full-scale reporting chain, with all the Norwegian municipalities and counties participating.

II.   THE REPORTING CHAIN OF IDUN AND KOSTRA

A.   IDUN

6.   Information sent to Statistics Norway from the different businesses does not contain personal data and is thus not considered sensitive in this respect. Statistics Norway thus has no duty towards the businesses or The Data Inspectorate[5], to encrypt the incoming data. However, as these data may contain information considered to be of value as business information or to the stock exchange, it has been decided to establish a

---

[3] IDUN - Informasjon og Data Utveksling for Næringslivet - Norwegian acronym meaning "Information and Data Exchange for Business and Commerce"
[4] KOSTRA - KOmune STat RApportering - Norwegian acronym meaning "Direct reporting of data from Local to Central Government Administrations"
[5] The Data Inspectorate - is an independent administrative body under the Norwegian Ministry of Labour and Government Administration, and its purpose is to protect persons from violation of their right to privacy through the processing of personal data.

secure channel where data from the companies can be encrypted. Thus the companies involved can feel more confident that privacy in their reporting is ensured.

7.   Today IDUN is in a pilot phase, and for that reason very few companies are reporting to Statistics Norway. There are 14 companies, and they all report only 2 forms to SN. In near future it will be thousands of businesses, which will report 10 forms and eventually all businesses will be reporting all forms available.
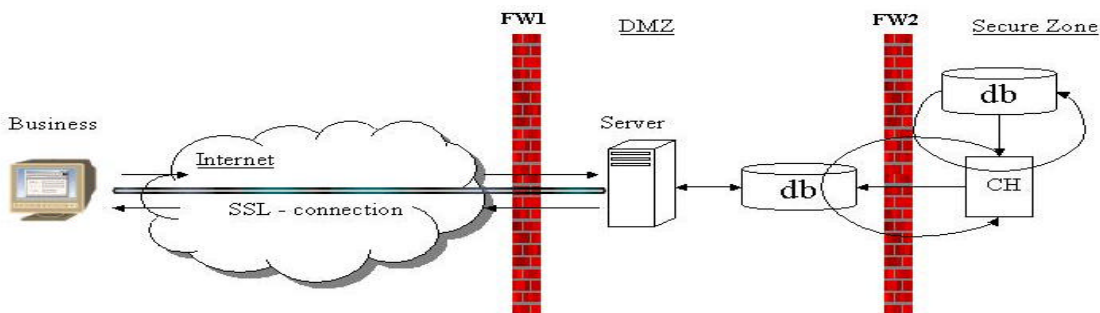
8.   As you can see from the figure above, all communication between the different companies and Statistics Norway will be onlined over the Internet. The businesses log on to a web site at SN, and when a server in Statistics Norway's DMZ[6] has confirmed the business' identity, that server establishes a SSL[7] connection to the specified business company. This server, is on SSB's DMZ, and is located behind the first firewall, it has a SSL certificate and the dataflow between SN and the businesses can be encrypted with 128-bits encryption in this secure channel. (For more information on SSL and security see Appendix)

9.   The "Custom House" (CH) works as an extension of the inner firewall, and is responsible for all dataflow between DMZ and the secure zone. It examines the data content, which is to be exchanged between the two zones. The dataflow is in XML[8] format.

10.   All communication between the security zone and DMZ is being initialised from the secure zone. The way it works is that the database in the secure zone polls the "Custom House", and sees if there is anything it shall pick up. The CH further polls against a database in DMZ, to see if there is something to get there.

11.   Thus, in a scenario where a company delivers a questionnaire form to SN it first sends the data to the database in DMZ via the secure channel established by the server in DMZ. The "Custom House" then polls the database in DMZ, and gets the form from it. It puts it in a queue for the database in the secure zone to pick it up. When the database in the secure zone has got the file it processes it and puts it back to the CH. The "Custom House" then puts this back through the firewall and into the DMZ database where the company can view it.

---

[6] DMZ - DeMilitarized Zone
[7] SSL - Secure Sockets Layer
[8] XML - eXtensible Markup Language

B.   KOSTRA

12.   In contrast to IDUN, which is an online report chain, KOSTRA is an offline based reporting chain. The municipalities work offline with their data, and when they are ready to ship the data, they are sent via email to Statistics Norway.

13.   KOSTRA is now in the final phase of the pilot, and all municipalities and counties send in 50 different forms to SN. Forms related to child-care, social-clients and family-protection offices contain sensitive information, and must be encrypted before they are sent to Statistics Norway.

14.   From the municipalities and the counties, there are now altogether four different kinds of data delivery: They are all delivered over SMTP[9], but two different data formats are used: EDIFACT[10], or XML. Further, both of these formats can be sent either as sensitive information, which must first be encrypted, or as non-sensitive information (not encrypted).

15.   The email address and subject field in the email determines the actions to be taken within SN. There are different email addresses depending on whether it is an EDIFACT message or an XML message that is delivered. If the message contains sensitive information, the questionnaire is encrypted, then the subject field in the email is marked with this, and the email is automatically delivered to a place different from that of messages that contain no sensitive information.

Reporting in KOSTRA

---
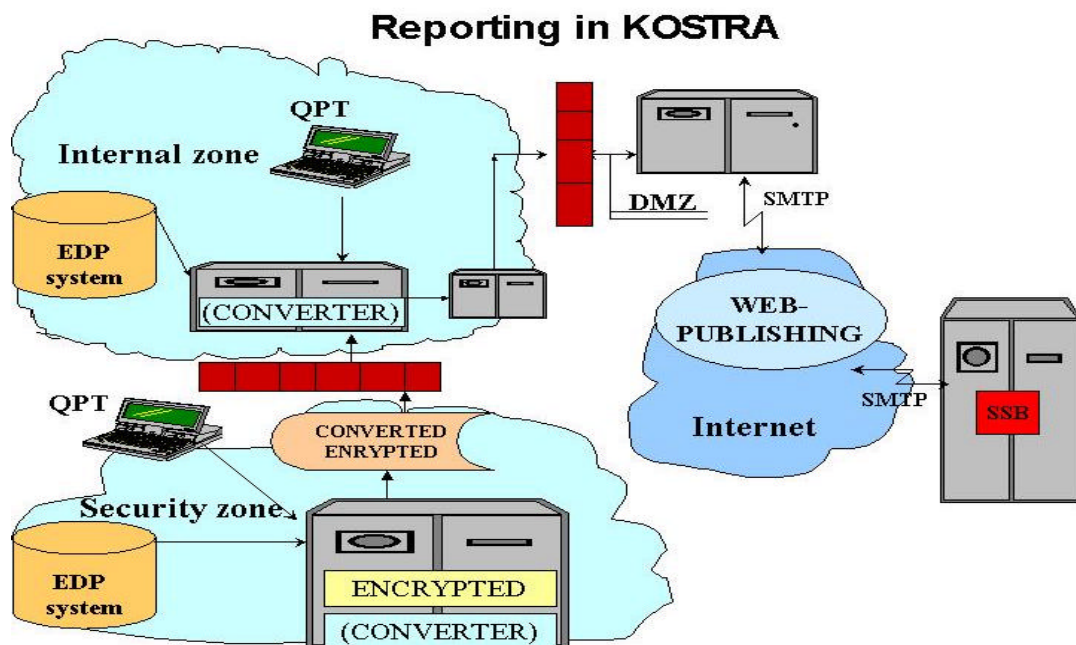
[9] SMTP - Simple Mail Transfer Protocol
[10] EDIFACT - Electronic Data Interchange For Administration, Commerce and Transport

16.   This is the general solution on the reporting chain in KOSTRA. This is a fully automatic solution that The Data Inspectorate has approved. There are also other solutions that are not fully automatic, requiring the municipalities to transfer their encrypted files from their secure zone to their internal zone, from where the files may be emailed to Statistics Norway.

17.   The reports are either extracted as a flat-file from the EDP-system or they are filled into a questionnaire form. Then the file is converted into EDIFACT or XML; depending on what EDI system they run. After the file has been converted, it is encrypted if it is person sensitive, and then sent over the Internet as an attachment to an email to Statistics Norway. SN receives the email and processes it according to the email address and subject field in the email. The respondent will get a receipt from Statistics Norway when the file is received and when SSB are finished processing the reports; the results are published on the web.

18.   For the reporting year 2001(which is reported in February 2002), all municipalities will be participating in the project. In the future all communication between Statistics Norway and the counties and municipalities will be based on the XML format. The reasons for this are:
 * The respondents may report cheaper and easier with XML format
 * The development of electronic questionnaires is cheaper and fasterWe benefit from our participation in the IQML[11] project (financed by the EU)
Other benefits are:
 * Flat-files from the local councils data systems are converted to XML easy and fast by using a special converting program and it is expected that vendors of this systems will export XML directly in the future XML Files from the electronic questionnaires are sent by e-mail direct to SN without converting the data; this eliminates the need for an EDI converter The XML-solution is probably going to be free as far as software and courses are concerned, the only requirement is IE 5.0 +
 * Limited adjustments (compared to the previous system) in the user interface regarding the questionnaire layout and data communication (still based on e-mail)XML usage in KOSTRA 2001:

19.   For the reporting year 2001, there will be a total of 22 municipalities that will report their forms using the XML format, they will report all their 24 questionnaires and all EDP[12]-systems. All 63 family protection offices will also report 2 questionnaires in XML format, and all county councils will report their 2 questionnaires in the XML format.

20.   Questionnaire Presentation Tool (QPT) : The distribution package (QPT) is sent to the respondent either on a CD-ROM, e-mail or downloaded via Internet. The CD-ROM contains a program that the participants must install, Microsoft's Internet Explorer 5.5, and all the questionnaires in printable pdf-format. When the municipalities or counties install the enclosed program, they will get access to their respectively questionnaire forms, and a program for converting existing flat-files into XML format. One requirement we have set for the municipalities or counties that whish to report with XML, is that they use Microsoft's Internet Explorer 5.0 or higher. This is because QPT use the Internet Explorer High Encryption Pack, which gives you 128-bit encryption, the highest level of protection possible for all your Internet communication, including credit card use and financial transactions. High encryption pack is included in the latest versions of the Internet Explorer browser (IE5.5+), and can be installed from IE5.0 and higher. Those municipalities or counties that does not have IE5.0 or IE5.5 can install it from the distribution package.

---

[11] IQML - Intelligent Questionnaire Markup Language
[12] EDP - Electronic Data Processing - this is where child-care, social-clients, and family-protection office data are located.

III. SOME EXPERIENCES / CONCLUSION

21.   KOSTRA is in its final phase and is therefore not undergoing major changes, except from the changes that must be made in order to use the XML format instead of EDIFACT for all communication in future. For the moment we have two working solutions that do the same thing. This dual technology is both unwanted and expensive, but is a solution we have to live with for the moment.

22.   XML have been well accepted and there are several reasons for that:
- XML is easier so understand than EDIFACT. EDIFACT is complicated and less intuitive than XML.
- XML is less expensive than the existing EDIFACT solutions. While XML is a free solution, EDIFACT costs each municipality around 15000 NOK.

23.   The result of XML beeing free is that two municipalities that refused to participate in the KOSTRA project, joined after getting the opportunity for a free XML solution.

24.   In KOSTRA we also have developed a way to develop forms directly from KOSTRA's meta database. These forms can be represented either as text, www or XML, and this solution is saving Statistisc Norway a lot of expenses in making the forms (approximately 400 000 NOK each year).

25.   Since the KOSTRA project now is testing the XML format, the IDUN project is waiting for the results on how the XML reporting in early 2002 will be received and assessed. If the results are good they want to use a similar approach in their online data report-chain. In the future SN will coordinate the two meta databases in KOSTRA and IDUN, and IDUN wants to use the same technique as KOSTRA in generating automatic questionnaire forms from the meta database.

26.   One "problem" that SN have encountered in KOSTRA is that some municipalities and counties are more or less reluctant to use a computer instead of traditional paper based questionnaire forms. This may be related to the fact that there is lack of computer knowledge in some of the smaller municipalities, and that the persons responsible for the data delivery to Statistics Norway then rather would prefer paper based delivery.

27.   To help the municipalities overcome their fear of using computers and electronic reporting as their way of reporting data to Statistics Norway, all municipalities get the oppurtunity for help and advise in using KOSTRA through small courses held by the companies that deliver the datasystem to the municipalities. For XML these courses is held by SN, which is responsible for the delivery of the XML system.

28.   Since 2002 is the first year XML is employed in KOSTRA, SN will meet many challenges. In the first round there are 22 municipalities and 19 counties that will participate in this new data delivery form. In addition data from all family protection offices will for the first time also be reported to Statistics Norway using XML. This means that even though this is the last year of the KOSTRA project (before it becomes a permanent report chain), there will be some need for further development and testing.

29.   Statistics Norway expects KOSTRA to become a stable solution in a couple of years. This is closely related to the fact that not all municipalities reporting this year have reported before, and from earlier experiences we know that municipalities that are new to KOSTRA need support. The forms can also be changed from year to year. These new changed forms need to be tested.

30.   Another problem we have encountered when dealing with converting flat files to XML, is that the size if the files may increase as much as 46 times as the original file. For municipalities that do not have an

approved solution by the Data Inspectorate, this means that they have to use a floppy disk or similar media to transport their files from their secure network to their DMZ. This becomes a problem if the file size exceeds 1.44 MB. One solution to this problem is of course for the municipalities to buy a cd-writer and use cd's to transport their files between their networks, or to split the files into several smaller files. Another solution may be that for the very big files, the respondents send in the flat files to SN for them to convert the flat files to their database format.

31.   There is no documentation on how much money or manpower KOSTRA is saving SN or the municipalities. In some aspects SN uses more money and manpower now than before KOSTRA started. This is mostly because before KOSTRA started, every ministry or research department that wanted some information from the municipalities asked the municipalities directly. As a result of this every municipalities gave away the same information several times to different ministries. Now all information is collected and sent to Statistics Norway via KOSTRA once, then SN delivers the information to the ministries that want it. This means that both the ministries and municipalities now have less workload than before, but Statistics Norway has a little more work to do, since SN now collects information for the ministries as well as for own purposes.

32.   With KOSTRA the municipalities get statistics that are more up to date than before. Now they get statistics one month after reporting it, in contrast to before when it could take over a year for the information to be processed. The municipalities can also look at other municipalities and compare statistics, which can be helpful for internal administration in the municipalities. The municipalities can look at other municipalities of similar size and see if they use more or less money/resources than the other municipalities.

33.   For IDUN to be a success, Statistics Norway needs to implement a user friendly, secure data report-chain solution, including adequate security for the respondents questionnaire forms during their internet travel towards SN, and in SN's way of storing the data. By adequate we mean enough security for the respondents to feel safe delivering data to Statistics Norway.

APPENDIX: SECURITY AND EXPLANATION OF TERMS

A.   Security

In computer terms data security is a conception of how to secure your data resources from being abused or illegally put out of function. In other words you can say that security is a set of safety rules and actions to protect our assets in a computer system. Mainly you can split security into three parts: confidentiality, integrity and availability.

Confidentiality: Means that your data should not be available for unauthorised persons. Confidential information is secret information, something that outsiders should not have access to. For instance: military information, private information, cryptography keys, and business secrets.

Integrity: Integrity, in terms of data and network security, is the assurance that information can only be accessed or modified by those authorized to do so. Measures taken to ensure integrity include controlling the physical environment of networked terminals and servers, restricting access to data, and maintaining rigorous authentication practices. Data integrity can also be threatened by environmental hazards, such as heat, dust, and electrical surges.

Availability: Means that a service meets certain demands to stability, so that the information you are looking for is available when you need it.

To achieve security we often use: authorisation, authentication, and cryptography.

Authorisation: Authorisation is the process of giving someone permission to do or have something. In multi-user computer systems, a system administrator defines for the system which users are allowed access to the system and what privileges of use.

Authentication: Authentication is the process of confirming a given identity. There are many ways to authenticate an identity, and we can split this in three categories:
- By using something you know; for example a password, PIN[13]…
- Token based, you use something you have; ID card, birth certificate, smart card…
- Biometric, by using something you are; for example fingerprint, voice recognition, retina and iris measurements, facial feature measurements…

You can also divide authentication into how it is performed:
- Manual; for example over the counter in the bank
- Electronic; for example login, cash dispenser
- Fully automatic; for example digital signature

In private and public computer networks (including the Internet), authentication is commonly done through the use of logon passwords. Knowledge of the password is assumed to guarantee that the user is authentic. The weakness in this system for transactions that are significant (such as the exchange of money) is that passwords can often be stolen, accidentally revealed, or forgotten. For this reason, Internet business and many other transactions require a more stringent authentication process. The use of digital certificates issued and verified by a Certificate Authority (CA) as part of a public key infrastructure is considered likely to become the standard way to perform authentication on the Internet.

---

[13] PIN - Personal Identification Number

Cryptography: Cryptography is the science of information security. The word is derived from the Greek *kryptos*, meaning hidden. In today's computer-centric world, cryptography is most often associated with scrambling plaintext into cipher text, a process called encryption, then back again, and known as decryption. Modern cryptography concerns itself with the following four objectives:

1) Confidentiality - the information cannot be understood by anyone for whom it was unintended

2) Integrity - the information cannot be altered in storage or transit between sender and intended receiver without the alteration being detected

3) Non-repudiation - the creator/sender of the information cannot deny at a later stage his or her intentions in the creation or transmission of the information

4) Authentication - the sender and receiver can confirm each other's identity and the origin/destination of the information

B.   SSL

The SSL (Secure Sockets Layer) protocol is a protocol used for managing the security of a message transmission on the Internet. The protocol is included in most of today's web-browsers and web-server products. SSL is using the RSA[14] public/private -key encryption system, which includes the use of a digital certificate.

The secure sockets layer protocol has become the universal standard on the web for authenticating sites for encrypting communication between users and web servers. Since SSL is a built in feature in most web-browsers, simply installing a digital certificate or server ID enables SSL capabilities.

SSL server authentication allows users to confirm a web server's identity. SSL - enabled client software, such as web browsers, can automatically check that a server's certificate and public ID are valid and have been issued by a certificate authority (CA), such as VeriSign, listed in the clients software's list of trusted CA's.

An encrypted SSL connection requires all information sent between a client and a server to be encrypted by the sending software and decrypted by the receiving software, protecting private information from interception on the Internet.

In addition, all data sent over an encrypted SSL connection is protected with a mechanism for detecting tampering that is, for automatically determining whether the data has been altered in transit. This means that the users can confidently send private data, such as credit card numbers, to a web site, thrusting that SSL keeps it private and confidential.

C.   PKI

A public key infrastructure (PKI) enables users of a basically insecure public network, such as the Internet to securely and privately exchange data and money through the use of a public and private cryptographic key pair that is obtained and shared through a trusted authority. The PKI provides a digital certificate that can identify an individual or organization, and directory services that can store and, when necessary, revoke the certificates.

---

[14] RSA - an encryption and authentication system developed by Ron Rivest, Adi Shamir ,Leonard Adleman

A PKI consists of:

- A certificate authority (CA) that issues and verifies digital certificate. A certificate includes the public key or information about the public key.
- A registration authority (RA) that acts as the verifier for the CA before a digital certificate is issued to a requestor.
- One or more directories where the certificate (with their public keys) are held.
- A certificate management system.

The public key infrastructure assumes the use of public key cryptography, which is the most common method on the Internet for authenticating a message sender or encrypting a message.

Digital certificate: A digital certificate is a certificate issued by an certification authority (CA) and contains your name, a serial number, expiration date, a copy of the certificate holders public key, and the digital signature form the certificate issuing authority, so that a recipient can verify that the certificate is real. Digital certificates can be kept in a register so users can collect each other's private keys.

Digital signature: A digital signature (not to be confused with a digital certificate) is an electronic signature that can be used to authenticate the identity of the sender of a message or the signer of a document, and possibly to ensure that the original content of the message or document that has been sent is unchanged. A digital signature can be used with any kind of message, whether it is encrypted or not, simply so that the receiver can be sure of the sender's identity and that the message arrived intact. A digital certificate contains the digital signature of the certificate-issuing authority so that anyone can verify that the certificate is real.

D.   EDI

EDI - Electronic Data Interchange is a standard format for exchanging business data. An EDI message contains a string of data elements, each of which represents a singular fact, such as a price, product model number, and so forth, separated by delimiter. The entire string is called a data segment. One or more data segments framed by a header and trailer from a transaction set, which is the EDI unit of transmission (equivalent to a message). A transaction set often consists of what would usually be contained in a typical business document or form.