WP. 11
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Neuchâtel, Switzerland, 5-7 October 2009)

Topic (ii): Editing near the source

COMPLEX METADATA DRIVEN GENERALIZED SYSTEM FOR STATISTICAL PROCESSING - CHALLENGES AND THREATS IN THE PROCESS OF DATA EDITING

Supporting Paper

Prepared by Metka Zaletel and Eva Belak, Statistical Office of the Republic of Slovenia

I. INTRODUCTION

1. The Statistical Office of the Republic of Slovenia (SORS) is developing a new infrastructure for data collection from primary sources. The ever growing requirements of the society and developments in other NSI's showed us that the introduction of new technologies into the processes is a prerequisite for achieving satisfying response rates, for successfully implementing the surveys and for optimising the cost. We expect that the results will lead to cost reduction at the reporting units and minimize the existing legacy of stove-pipe processes at our NSI.

2. Along with the introduction of new data inflow channels and on-line tracking, the change is also needed in the internal processes and organization with the introduction of a new input data warehouse which should be fast, have the possibility of tracking changes, be quickly adaptable to the changes in the environment and enable quick responses to the demands from subject-matter statisticians to the ever growing personalisation of the contacts with reporting units.

3. It is also obvious that due to the mentioned new procedures the statistical office should strive to offer support for the reporting units. In this way we are planning to introduce a new service (help desk) for internal and external users of the new process.

4. The standardisation of the questionnaires at SORS comprises:
- Visual layout of the questionnaires (design of the thematic modules, design of the questions, navigation throughout the questionnaire, standardized questions (paradata) regarding the completion of the questionnaire - time, remarks, information about the person who completed the questionnaire, etc.);
- Standardized forms for the communication with responding units (enterprise): advanced letters, reminders, etc.;
- A common database of the variables and a database of the questions.

5. The final goal of the standardization is to set up an application which will make the procedures at SORS uniform, including:
- Registration or selection of the variables and the questions which will be printed on the questionnaire;
- Automatic preparation (with minimum “manual” design) of the questionnaires.

6. In the paper, first the features of the new systems are explained; later, the case study is presented.

II. NEW METADATA SYSTEM FOR THE PREPARATION OF THE QUESTIONNAIRES

7. Application for preparation and management of variables, questions and questionnaires is part of a larger system ISIS and is called “Survey Design Module” (SDM). It is a metadata supported system with two important features: other SORS metadata systems are integrated (register of statistical surveys, detailed business plan and classification server), and secondly, questionnaires and questions are parameterised – this means that reference periods and other parameters are visible on the questionnaire when the questionnaire is linked to the instance (concrete realisation of a survey questionnaire).

8. A prerequisite for the application is that visual and technical elements of the questionnaire are standardised. The standardisation is described in the next chapter. In reality that means that the questionnaire is divided into thematic modules, questions, and sub-questions; and for different elements in the questionnaire the size of letters, boxes for ticking the answers, location of notes and remarks, etc., are defined.

9. There are several steps to prepare the questionnaire in ISIS. Firstly, some activities have to be taken into account even before using the system:

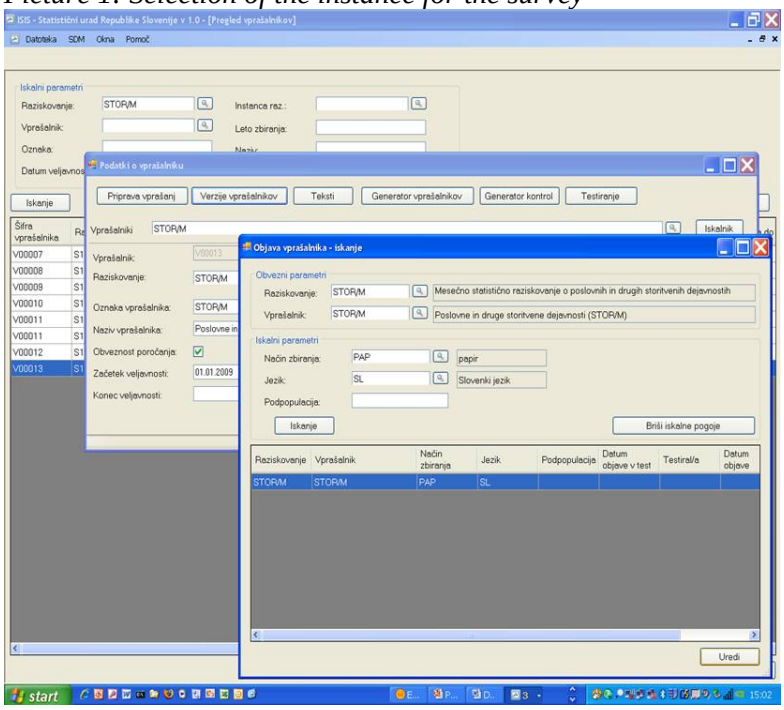
- The subject-matter statistician should consider newly adopted standards when preparing the questionnaire. The questionnaire is firstly prepared in Word and is then cognitively tested (focus groups, think aloud protocols, etc.).
- Complex tables in existing questionnaires should be simplified, since SDM does not support complex tables and also such tables are not very intuitive to the respondents.
- After the questionnaire is tested, the subject-matter statistician prepares a dataset of questions and variables if he or she will import the metadata with the help of a dataset.

10. After that, ten steps of preparing the questionnaire in the ISIS system follow:
(i) Registration of the new questionnaire: open the new questionnaire, fill out compulsory information and save. When the data are saved, ID of the questionnaire is defined.
(ii) Registration of the questions and variables: there are three possibilities for the registration:

- Metadata of the questionnaire are prepared in Excel and then the Excel file is imported into the application (csv format);
- Questions of the existing questionnaire are copied into the new questionnaire;
- Manual entry of the questions into the questionnaire.

(iii) Selection of the instance for the survey.

Picture 1: Selection of the instance for the survey



(iv) Definition of the logical controls.
(v) Formal (not cognitive) testing of the questionnaire (Publish in test).
(vi) The subject-matter unit delivers the tested questionnaire to the print unit to finalize the questionnaire.
(vii) Finalisation of the questionnaire (print unit).
(viii) The questionnaire is finalised and confirmed by the subject-matter unit.
(ix) Publish the questionnaire into the production.
(x) Merging variable information on the questionnaire: personalization of the questionnaire: barcode; pre-print of the information if it is envisaged.

III. STANDARDISATION OF THE QUESTIONNAIRES

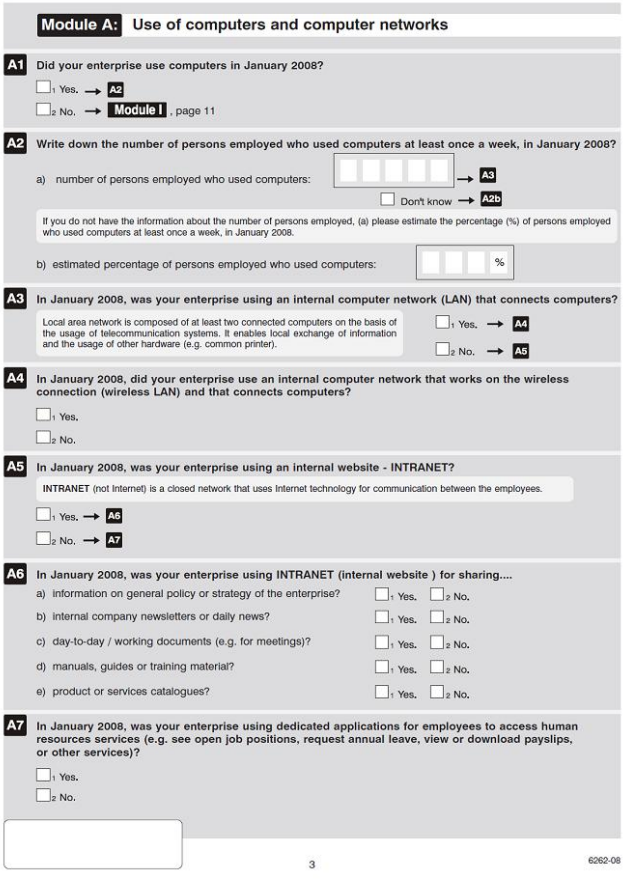
11. In 2008, standardised elements of the questionnaire were defined for business surveys. At SORS we took as a model the design of the questionnaires of Statistics New Zealand, which we thought was a good example of the visual presentation of the questionnaire. Standardisation usually comprises at least three different points of view: (a) visual and technical standardisation, (b) content standardisation, and finally, (c) standardisation of materials that accompany the questionnaire (letters, reminders, brochures).

12. Paper questionnaires are adapted to optical scanning; naturally, visual and technical standards of scanning software were strictly followed. The following elements are standardised:

- Size and font of the letters,
- Design of modules, questions and sub-questions,
- Guidelines for individual questions (module, question, sub-questions, etc.),
- Size and shape of answer boxes for recording of text answers,
- Size and shape of answer boxes for recording monetary values,
- Size and shape of answer boxes for ticking the answers.

The result of standardisation and the division of the questionnaire into thematic modules can be seen on the below pictures (case study – Survey on the use of information and communication technology in enterprises):

Picture 2: Standardised questionnaire – case study



Module A: Use of computers and computer networks

A1 Did your enterprise use computers in January 2008?
☐ 1 Yes. → A2
☐ 2 No. → Module I, page 11

A2 Write down the number of persons employed who used computers at least once a week, in January 2008?
a) number of persons employed who used computers: → A3
☐ Don't know → A2b
If you do not have the information about the number of persons employed, (a) please estimate the percentage (%) of persons employed who used computers at least once a week, in January 2008.
b) estimated percentage of persons employed who used computers: %

A3 In January 2008, was your enterprise using an internal computer network (LAN) that connects computers?
Local area network is composed of at least two connected computers on the basis of the usage of telecommunication systems. It enables local exchange of information and the usage of other hardware (e.g. common printer).
☐ 1 Yes. → A4
☐ 2 No. → A5

A4 In January 2008, did your enterprise use an internal computer network that works on the wireless connection (wireless LAN) and that connects computers?
☐ 1 Yes.
☐ 2 No.

A5 In January 2008, was your enterprise using an internal website - INTRANET?
INTRANET (not Internet) is a closed network that uses Internet technology for communication between the employees.
☐ 1 Yes. → A6
☐ 2 No. → A7

A6 In January 2008, was your enterprise using INTRANET (internal website) for sharing....
a) information on general policy or strategy of the enterprise? ☐ 1 Yes. ☐ 2 No.
b) internal company newsletters or daily news? ☐ 1 Yes. ☐ 2 No.
c) day-to-day / working documents (e.g. for meetings)? ☐ 1 Yes. ☐ 2 No.
d) manuals, guides or training material? ☐ 1 Yes. ☐ 2 No.
e) product or services catalogues? ☐ 1 Yes. ☐ 2 No.

A7 In January 2008, was your enterprise using dedicated applications for employees to access human resources services (e.g. see open job positions, request annual leave, view or download payslips, or other services)?
☐ 1 Yes.
☐ 2 No.

3 6262-08

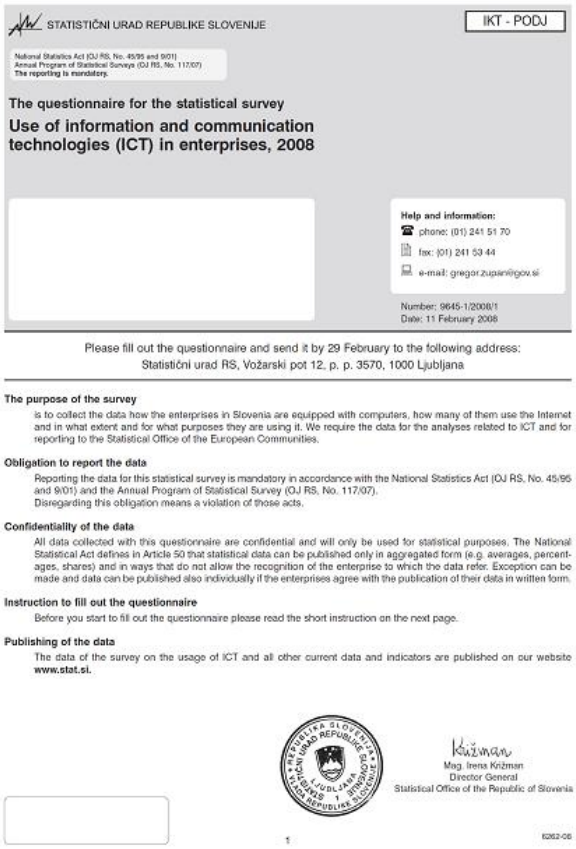
13. Then, the colour of the questionnaires has been uniformly defined: the paper questionnaires are black and white, with grey raster areas, which divides questions and modules of the questions. The decision on gray colour scale was made almost purely on the cost basis since the expenses are approximately 10 times lower when compared to those of colour digital printing.
Electronic questionnaires are in colour and should be printable since enterprises should be able to print blank or completed questionnaires.

14. At the same time, the first page of the questionnaire, the second page ("instruction page") and the last page were standardised too. On the last page, some paradata are collected:

- Time for completing the questionnaire,
- Information concerning time for completing the questionnaire and preparing the data for the questionnaire needed for the calculation of the response burden,
- Contact information: name/surname of the person who filled out the questionnaire, function in the enterprise, phone, fax number, e-mail address, signature, date of completing the questionnaire.

These data are also scanned and stored into the database. The question on frequency of questions on time for completing the questionnaire still has to be answered. The idea is that this question is included every few years or when the questionnaire is revised or new questions are introduced into the questionnaire.

Picture 3: The first page of the questionnaire



STATISTIČNI URAD REPUBLIKE SLOVENIJE

IKT - PODJ

National Statistics Act (OJ RS, No. 45/95 and 9/01)
Annual Program of Statistical Surveys (OJ RS, No. 117/07)
The reporting is mandatory.

The questionnaire for the statistical survey

Use of information and communication technologies (ICT) in enterprises, 2008

Help and information:
phone: (01) 241 51 70
fax: (01) 241 53 44
e-mail: gregor.zupan@gov.si

Number: 9645-1/2008/1
Date: 11 February 2008

Please fill out the questionnaire and send it by 29 February to the following address:
Statistični urad RS, Vožarski pot 12, p. p. 3570, 1000 Ljubljana

The purpose of the survey

is to collect the data how the enterprises in Slovenia are equipped with computers, how many of them use the Internet and in what extent and for what purposes they are using it. We require the data for the analyses related to ICT and for reporting to the Statistical Office of the European Communities.

Obligation to report the data

Reporting the data for this statistical survey is mandatory in accordance with the National Statistics Act (OJ RS, No. 45/95 and 9/01) and the Annual Program of Statistical Survey (OJ RS, No. 117/07).
Disregarding this obligation means a violation of those acts.

Confidentiality of the data

All data collected with this questionnaire are confidential and will only be used for statistical purposes. The National Statistical Act defines in Article 50 that statistical data can be published only in aggregated form (e.g. averages, percentages, shares) and in ways that do not allow the recognition of the enterprise to which the data refer. Exception can be made and data can be published also individually if the enterprises agree with the publication of their data in written form.

Instruction to fill out the questionnaire

Before you start to fill out the questionnaire please read the short instruction on the next page.

Publishing of the data

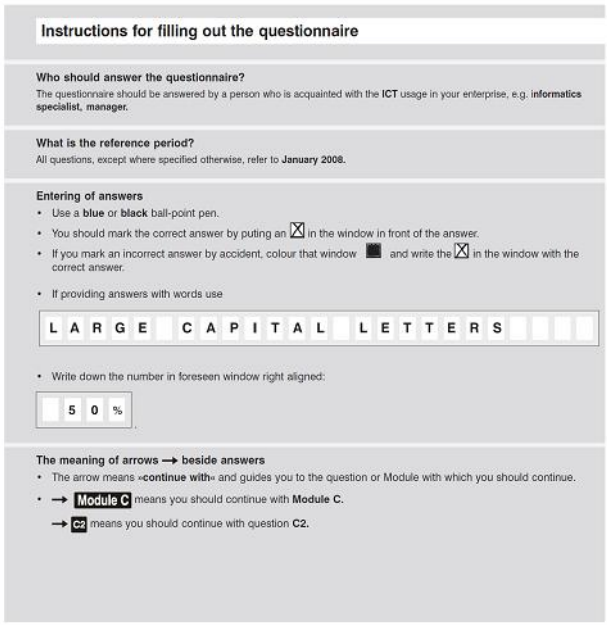
The data of the survey on the usage of ICT and all other current data and indicators are published on our website **www.stat.si.**

Mag. Irena Križman
Director General
Statistical Office of the Republic of Slovenia

1

6262-06

Picture 4: The second page of the questionnaire



Instructions for filling out the questionnaire

Who should answer the questionnaire?
The questionnaire should be answered by a person who is acquainted with the **ICT** usage in your enterprise, e.g. **informatics specialist, manager.**

What is the reference period?
All questions, except where specified otherwise, refer to **January 2008.**

Entering of answers
- Use a **blue** or **black** ball-point pen.
- You should mark the correct answer by puting an ☒ in the window in front of the answer.
- If you mark an incorrect answer by accident, colour that window ■ and write the ☒ in the window with the correct answer.
- If providing answers with words use

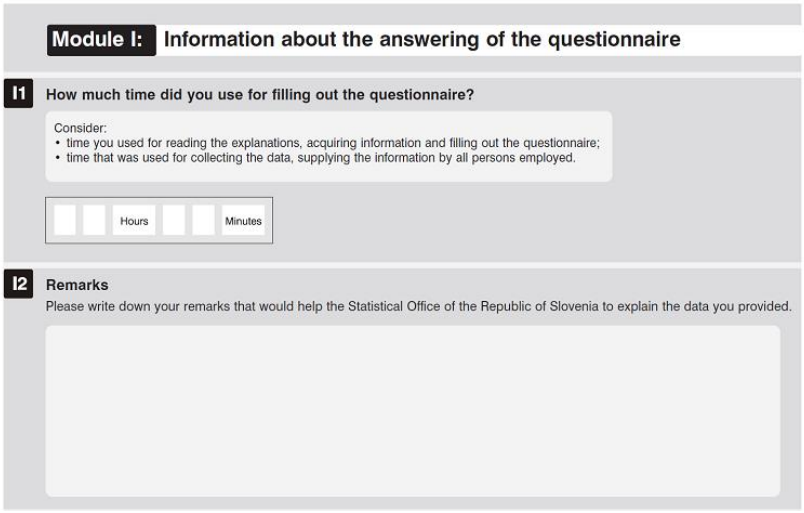
L A R G E C A P I T A L L E T T E R S

- Write down the number in foreseen window right aligned:

5 0 %

The meaning of arrows → beside answers
- The arrow means »**continue with**« and guides you to the question or Module with which you should continue.
- → **Module C** means you should continue with **Module C.**
  → **C2** means you should continue with question **C2.**

Picture 5: The last page of the questionnaire



Module I: Information about the answering of the questionnaire

I1 **How much time did you use for filling out the questionnaire?**

Consider:
- time you used for reading the explanations, acquiring information and filling out the questionnaire;
- time that was used for collecting the data, supplying the information by all persons employed.

Hours Minutes

I2 **Remarks**

Please write down your remarks that would help the Statistical Office of the Republic of Slovenia to explain the data you provided.

IV. CASE STUDY

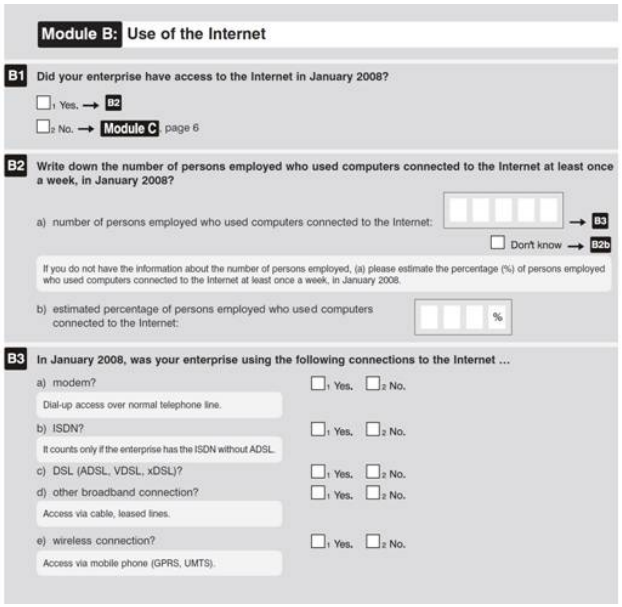
15. The survey on the use of information and communication technology has been taken as a case study for different reasons. Maybe the most important reason is that different types of variables are collected, therefore standardisation could have been performed and tested in probably the most intensive way. Basic characteristics of the survey are as follows:

- Annual survey on the usage of information and communication technologies in enterprises;
- SORS conducted the survey for the fifth time, the methodology is harmonised in the EU;
- The sample was selected form the Business Register; sample size was 2034;
- Mode of Data Collection: self–administrated paper questionnaire, no return envelope enclosed;
- 10% of the enterprises send their questionnaires via e-mail (enterprise could opt for improvised Excel questionnaire);
- Non-response: approximately 10% of the enterprises did not respond, although provision of information for the survey is compulsory for legal entities (businesses, agricultural holdings, etc.);
- Two remainders were sent out: first 3 days after due date and second 17 days after due date. A fine is also mentioned in the reminder;

– Telephone follow-up: controllers contacted enterprises to remind them to complete and return the questionnaire.

16. Two datasets of ICT 2008 were analysed: initial dataset (scanned data) and final dataset. If data in the initial dataset were different from data in the final dataset, then those data were flagged.

Picture 6: Number of edits for some questions in the questionnaire



Module B: Use of the Internet

B1 Did your enterprise have access to the Internet in January 2008?

☐1 Yes. → B2

☐2 No. → Module C, page 6

B2 Write down the number of persons employed who used computers connected to the Internet at least once a week, in January 2008?

a) number of persons employed who used computers connected to the Internet: → B3

☐ Don't know → B2b

If you do not have the information about the number of persons employed, (a) please estimate the percentage (%) of persons employed who used computers connected to the Internet at least once a week, in January 2008.

b) estimated percentage of persons employed who used computers connected to the Internet: %

B3 In January 2008, was your enterprise using the following connections to the Internet ...

a) modem? ☐1 Yes. ☐2 No.
Dial-up access over normal telephone line.

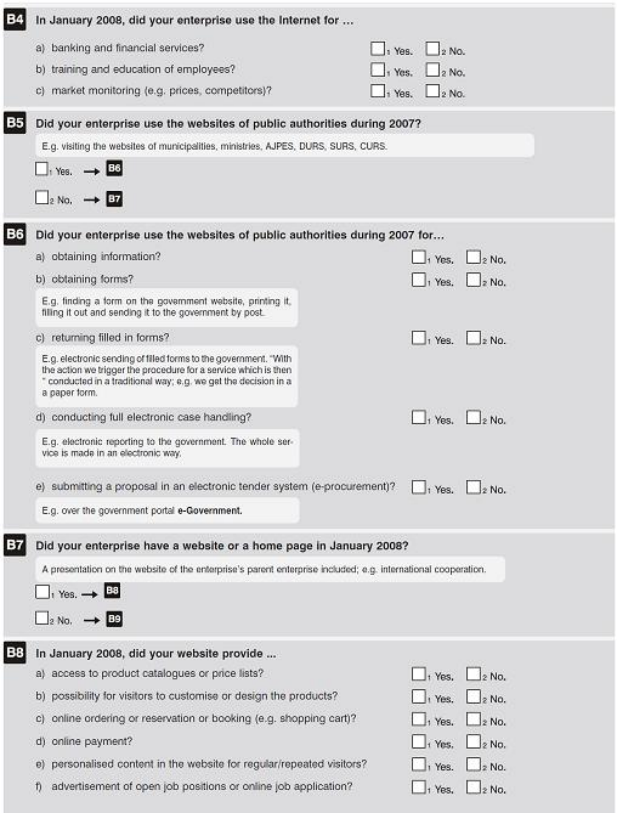
b) ISDN? ☐1 Yes. ☐2 No.
It counts only if the enterprise has the ISDN without ADSL.

c) DSL (ADSL, VDSL, xDSL)? ☐1 Yes. ☐2 No.

d) other broadband connection? ☐1 Yes. ☐2 No.
Access via cable, leased lines.

e) wireless connection? ☐1 Yes. ☐2 No.
Access via mobile phone (GPRS, UMTS).

Question	# of edits	% of edits
B1	44	2.7
B2a	66	4.3
B2b	47	3.0
B3a	466	30.0
B3b	425	27.4
B3c	247	15.9
B3d	476	30.7
B3e	449	28.9



B4 In January 2008, did your enterprise use the Internet for ...

a) banking and financial services? ☐1 Yes. ☐2 No.

b) training and education of employees? ☐1 Yes. ☐2 No.

c) market monitoring (e.g. prices, competitors)? ☐1 Yes. ☐2 No.

B5 Did your enterprise use the websites of public authorities during 2007?
E.g. visiting the websites of municipalities, ministries, AJPES, DURS, SURS, CURS.

☐1 Yes. → B6

☐2 No. → B7

B6 Did your enterprise use the websites of public authorities during 2007 for...

a) obtaining information? ☐1 Yes. ☐2 No.

b) obtaining forms? ☐1 Yes. ☐2 No.
E.g. finding a form on the government website, printing it, filling it out and sending it to the government by post.

c) returning filled in forms? ☐1 Yes. ☐2 No.
E.g. electronic sending of filled forms to the government. "With the action we trigger the procedure for a service which is then " conducted in a traditional way; e.g. we get the decision in a a paper form.

d) conducting full electronic case handling? ☐1 Yes. ☐2 No.
E.g. electronic reporting to the government. The whole service is made in an electronic way.

e) submitting a proposal in an electronic tender system (e-procurement)? ☐1 Yes. ☐2 No.
E.g. over the government portal **e-Government.**

B7 Did your enterprise have a website or a home page in January 2008?
A presentation on the website of the enterprise's parent enterprise included; e.g. international cooperation.

☐1 Yes. → B8

☐2 No. → B9

B8 In January 2008, did your website provide ...

a) access to product catalogues or price lists? ☐1 Yes. ☐2 No.

b) possibility for visitors to customise or design the products? ☐1 Yes. ☐2 No.

c) online ordering or reservation or booking (e.g. shopping cart)? ☐1 Yes. ☐2 No.

d) online payment? ☐1 Yes. ☐2 No.

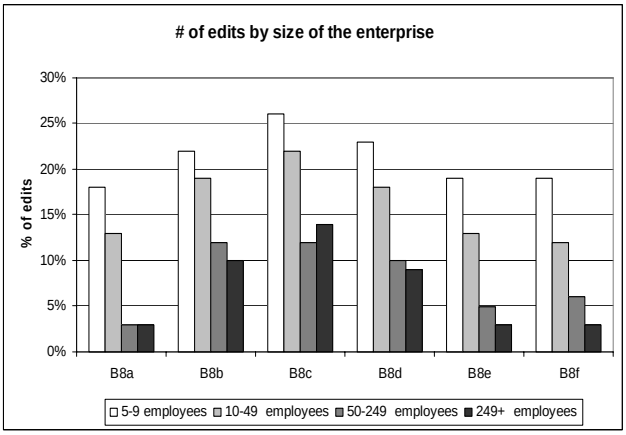
e) personalised content in the website for regular/repeated visitors? ☐1 Yes. ☐2 No.

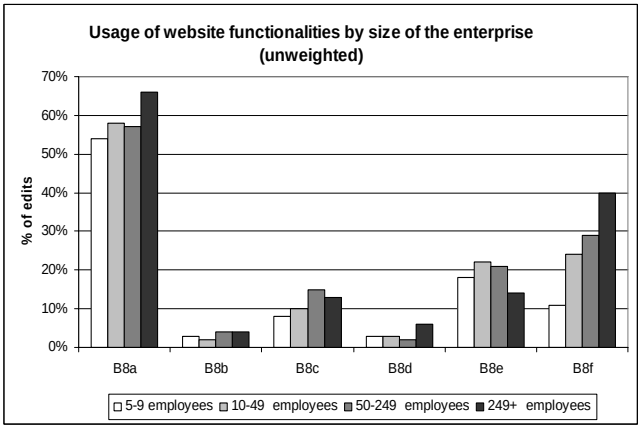
f) advertisement of open job positions or online job application? ☐1 Yes. ☐2 No.

Question	# of edits	% of edits
B4a	47	3.0
B4b	195	12.6
B4c	133	8.6
B5	94	6.1
B6a	126	9.2
B6b	204	14.9
B6c	246	18.0
B6d	159	11.6
B6e	180	13.2
B7	58	4.2
B8a	210	20.0
B8b	304	29.0
B8c	354	33.7
B8d	297	28.3
B8e	222	21.2
B8f	216	20.6
B9	51	4.9

17. Number of edits does significantly differ according to the size of the enterprise, which can be easily seen from the pictures below and is, of course, affected also by the content of the question.

Picture 7: Number of edits for some questions according to the size of enterprise





V. LESSONS LEARNT

18. After the results of the case study, there are, of course, several questions to discuss. The main question is, of course, the impact of above described activities on the quality of completed questionnaires.

(a) The impact of standardisation of questionnaires is expected to be seen in the long run. No matter which questionnaire is received from SORS and regardless of the mode of data collection (paper/e-reporting), the rules are clear and well known in advance. During the first and second year of implementation, maybe even more difficulties are expected and also have been experienced because of changes of questionnaires and rules that reporting units were used to fill in. In general, less edits at SORS are expected, but this hypothesis could be tested in a few years only.

(b) The impact of implementation of a new metadata driven system: rules are clearer also internally (within SORS). Again, the effect could be tested in a few years only, since the implementation of a new system is time consuming and very burdensome because of the necessary transfer of all questions and questionnaires into the system. Maybe the most important step forward is central supervision of all questionnaires, which enables us to ensure that similar questions and blocks of questions are presented in a similar way.

(c) The impact of the size of the enterprise: there were much fewer edits on the questionnaires received from large enterprises. There are several reasons to get this outcome:

- Larger enterprises can appoint more qualified people for filling out the questionnaire (IT specialist, someone who is acquainted with the ICT);
- And also, larger enterprises might more often respond with “Yes” due to more spread usage of the ICT in their enterprises. The effect of this fact could be seen especially within the battery of yes/no questions, where large enterprises are more likely to respond “yes”.

The challenge is how to improve the quality of completed questionnaire in small and medium-sized enterprises. It has to be emphasized that the case study was performed only with paper questionnaires; the problems of battery yes/no questions are eliminated in the e-reporting system. But in general, it is very difficult to overcome the problems of under-qualified employees/reporters. We could help them with easy questionnaire and clear rules, but lack of knowledge cannot be compensated.