INVOLVING INNOVATION IN DATA COLLECTION PROCESS IN PRODUCING OFFICIAL STATISTICS

ALMA KONDI, INSTITUTE OF STATISTICS
akondi@instat.gov.al

HELDA CURMA, INSTITUTE OF STATISTICS
hmitre@instat.gov.al

Abstract

Albanian Institute of Statistics (INSTAT) vision is to provide reliable and comparable data, adapting methodologies and adding a list of statistical indicators. Based on this vision the use of technology in different statistical processes has been increased during the last years, starting from data collection, editing, estimation, tabulation and dissemination. INSTAT has made progress in its use of technology (CAPI method, web-based questionnaires, OCR, R, PxWeb).

This article review the situation in 2016 regarding use of innovative technologies, improvements made in comparison with traditional methods of data collection, processing and publishing, and looks forward to the next challenges of INSTAT. The method used in this article is a deep analysis of cost, response rate, data capture and quality of the data in different methods of data collection.

The use of innovation can be made by the statistical offices to reduce cost and response burden, to maximize the quality of the data collected and improves timeliness.

KEY WORDS:

INSTAT; Data Collection; Technology; Statistical Quality

1. INTRODUCTION

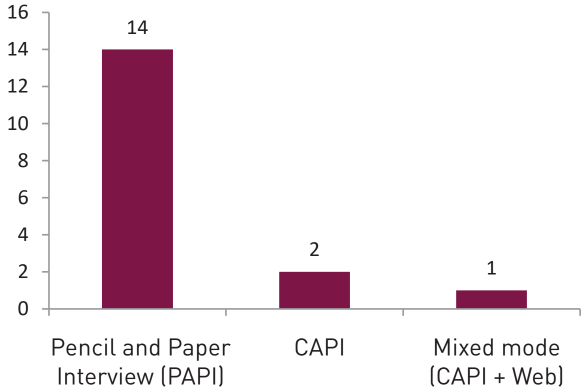
One of the most important statistical office's challenges is increasing quality and timeliness in a more cost effective way. The process of increasing high quality data for NSOs' (National Statistical Offices), is embracing the best practices in the collection, reducing cost and response burden. Costs are very important, especially for statistical offices with low budgets. The new technological changes can affect every statistical process. The rapid changes in technology have created new opportunities for improving timeliness. Costs of innovative technologies, such as tablets or other mobile devices are declining, making possible the use on the data collection process through the Computer-Assisted Personal Interviewing (CAPI) method.

Statistics Canada analyzes in the Methodology Report (Statistics Canada, 2010) the advantages of using technology in statistical production. Technology can monitor and control the quality of the data from data collection, data capture and also in the process of editing and imputation. Also, automated skip patterns can be used to resolve immediate problems, resulting in less follow-up and reduce the burden of the interviewer. It is easier to implement quality control of the interviewing process and generate management reports on the status of the interviews e.g., response rate, number of interviews completed, and length of time per interview, etc. On the other hand, the uses of technology require investment of buying or renew. It also requires computer experts to develop programs for the statistical production.

INSTAT's work is based on the five year Program of Official Statistics, which is the basic document that provides the production of statistical data by the National Statistical System, measuring different statistical process of economic, social and environmental situation in the Republic of Albania. This program works in accordance with the statistical principles provided in the Statistical Law but also follows the general principles of the European Statistics code of Practice.

During 2016, seventeen surveys were conducted by INSTAT. In Figure 1 are shown the number of surveys by the data collection methods for the year 2016.

Figure 1: Number of surveys by data collection methods



Source: INSTAT Program for 2016, calculation of the authors

Traditionally INSTAT has been using Personal Assisted Paper Interview (PAPI) as a data collection method. Nowadays it is still the most used method, as shown in Figure 1; around 82 percent of surveys are using the PAPI method to collect information, due to several factors such as generally high response rate, poor system of addresses, user friendly and often because it is cheaper.

For the first time CAPI was used in 2008-09 Albania Demographic and Health Survey (ADHS). An innovative aspect of the 2008-09 ADHS was the use of Personal Digital Assistants (PDAs) for the data collection, rather than paper questionnaires. The questionnaires were programmed in PDAs using the software package Census and Survey Processing (CSPro). Full survey and data management, range, skip and consistency checking were built into the data capture system. Paper questionnaires were available to interviewers in case of equipment failure. In 2016 the CAPI methods were used by 2 surveys.

The online data collection (WEB) was first used in 2013, as a way to lower costs, avoiding the use of interviewers, and reducing respondent's burden. For this purpose, an open source solution, Limesurvey Server was used. Various forms have been prepared for the collection of information, for example Short Term Statistics Survey, survey of Research Development for Public Sector, etc. This method was used in 2016 in one survey but was also used with CAPI method to provide more information.

2. COMPARISON OF DATA COLLECTION METHODS

This section will describe the different data collection methods, their advantages and disadvantages in different surveys conducted by INSTAT, in various perspectives such as cost, response rate, response burden, quality indicators, etc. These aspects are addressed in this article by analyzing and comparing them in the following surveys:

- Labour Force Survey (LFS)
- Short term Statistics Survey (STS)
- Information Communication and Technology Survey (ICT)

LFS was conducted for the first time in 2007. During 2007-2011 LFS has been carried out by INSTAT on annual basis through direct personal interviews, contacting households at their dwelling. In 2011, for the first time, LFS started using CAPI method. The optimized method (in our case CAPI method), allows us to have more accurate data because logical controls have been directly done in the application of the laptops. These controls reduce non-sampling errors (e.g. filter questions are skipped automatically and not manually, as it happens in paper questionnaires). For example, if a person is younger than 15 years old, the questionnaire regarding employment status is skipped automatically.

STS Survey is a quarterly survey which is addressed to the enterprises. From 2003-2013 the PAPI method has been used to collect the information. Since 2013 efforts has been made to use a mixed method approach: PAPI + online questionnaires. In 2014, 91% of companies, with 10 + employed, have had internet access. The high internet penetration rate, especially among big companies, is the reason why the web approach is being used for this survey. PAPI is applied as the base method, while web forms are being used for the big companies only (approximately 300).

ICT Survey was conducted in 2015 for the first time, using tablets (CAPI) to collect the information. This survey is focused on the availability of information and communication technologies (ICTs), and their use by enterprises.

2.1. COST

There is a growing demand on statistical data from users, not only from policy makers but from other type of users as well. Every statistical agency needs to make decisions between producing statistical data and the costs to do it. Collecting such data is costly, especially in countries with poor system of addresses and infrastructure. Costs affect every statistical process. In this article data collection costs are taken into consideration, as generally this process is the most expensive one. The budget items taken into account to analyze cost per questionnaire (in ALL) are shown in the Table 1.

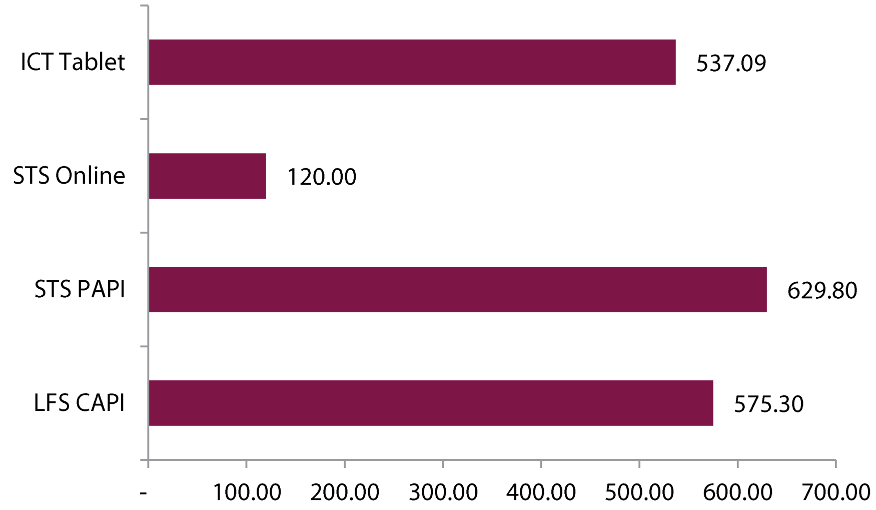
Table 1: Surveys budget items taken into account

	LFS CAPI	STS PAPI	STS Online	ICT Tablet
Printing of questionnaires	NA	Value	NA	NA
Printing of additional materials	Value	Value	NA	Value
Interviewer	Value	Value	NA	Value
Interviewer No	Value	Value	NA	Value
Controller	Value	Value	NA	NA
Consumables	Value	Value	NA	Value
Per diem for interviewers	NA	Value	NA	Value
Data Entry Operator	NA	Value	NA	NA
Internet Cards	Value	NA	NA	Value
Postal Service	NA	NA	Value	NA
Total	Value	Value	Value	Value

Source: Extracted by the authors based on INSTAT information

Figure 2 shows the surveys costs per questionnaire. As it is expected, the online approach by using web forms is the cheapest one. This approach is being used until now only for STS big companies, knowing for sure that they have in place the infrastructure to respond the questionnaire and also the statistical office has the possibility to send them their authentication credentials by post or email. Costs for CAPI using laptops or tablets are almost the same, excluding the initial costs for equipment, which can be very different, depending from the technologies used; although it is decreasing very rapidly. The PAPI method, which is still the most used method at INSTAT has the highest cost, from the other collection method.

Figure 2: Costs per questionnaire according different data collection approaches and surveys (in ALL)



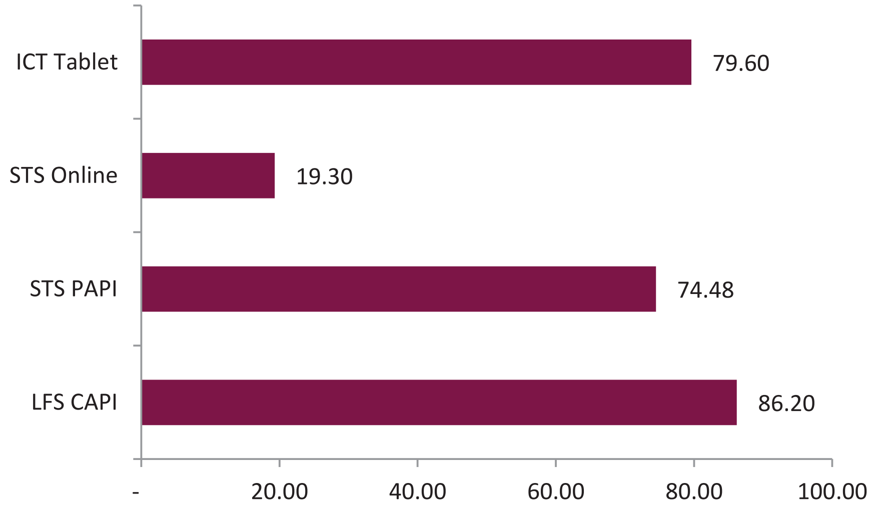
Source: Authors' calculations

2.2. RESPONSE RATE

Non-response is a common feature of sample surveys. A part of the sampled elements cannot be contacted, refuses participation or does not participate in the survey for other reasons. (Fannie Cobben, 2009)

In Albania, to complete voluntary a questionnaire is relatively a new culture. Generally, people cooperate with interviewers to fulfil the questionnaires. Still, it had become difficult to find respondents at home during daytime hours as the lifestyle has changed. The population has become more mobile and people are not at their usual place of residence for extended periods. All these changes can affect the response rate. For this purpose it is calculated the response rate shown in the Figure 3. The response rate is calculated by dividing the number of completed questionnaires by the total number of eligible units in the sample chosen.

Figure 3: Response rates according different data collection approaches and surveys (in %)



Source: Authors' calculations

There is a specific situation in Albania, because the system of addresses has serious weaknesses. This makes it very difficult to mail questionnaires or authentication credentials for online forms. For household surveys, INSTAT is using its own GIS system, designed for Population and Housing Census 2011 (PHC). The response rate for LFS is higher due to this fact. The response rate for LFS is the annual average of 2014, for STS PAPI is the annual average of 2013, for STS online is the rate for fourth quarter 2015, while for ICT tablet is the rate of 2015.

Bradburn (1978) suggested that the definition of respondent burden should include four elements: interview length; required respondent effort; respondent stress; and the frequency of being interviewed. The effort required of respondents could refer to the cognitive challenges of a task. The smallest response rate is for the online approach; 19.3%. Even with the online surveys, observation units can respond when it is appropriate for them, the burden on respondents is higher, as they need to understand, complete, keep notes, and calculate questions.

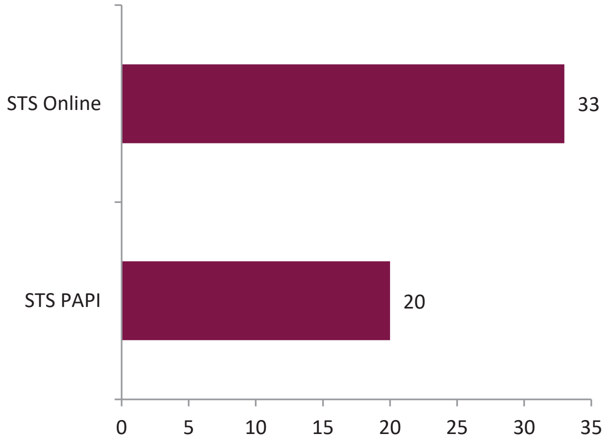
Traditionally, using interviewers for data collection has resulted the most efficient way regarding response rates; approximately 80%. The respondents trust interviewers and collaborate with them better, resulting in higher response rates. Interviewers are equipped with identification badges and have with them a formal paper explaining the scope of the survey and the data confidentiality aspects.

2.3. RESPONSE BURDEN

The measurement of response burden is done by taking into account duration of interview, for the same survey, in different data collection methods. STS survey has been tested in two different methods; PAPI and online. It is used the same questionnaire for both methods. On the online form there have been added additional notes, to explain what the interviewer does in the face to face method. The online form was directed to the same individual as per PAPI method (during PAPI method contact points have been decided previously).

Obviously the online method has a higher burden on the observation unit (in our case companies). This relates mostly to the fact that the respondent needs more time to understand the questions.

Figure 4: Average time spent on PAPI and Online method (in minutes)



Source: Authors' calculations

2.4. DATA CAPTURE

High dynamic developments in Information Technology (IT) and rapid advances in technology changes are affecting the way of work for the statistical offices. INSTAT's everyday challenge is to identify trends and find the most suitable solutions to produce statistics in a more cost effective way. Almost every survey carried out by INSTAT is using the face to face interview. To facilitate the data capturing process, a scanning system was installed in October 2009 to prepare the coming work for scanning censuses and large surveys questionnaires.

INSTAT's achievements regarding data capture are:

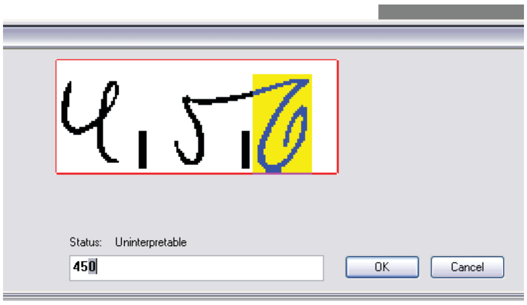
- In 2011 scan and capture all data from the Enterprise Census
- In 2011-2012 scan and capture all data from the PHC
- In 2013 scan and capture all data from the Agriculture Census
- From 2010 – till present scan and capture periodically surveys

The Intelligent Character Recognition (ICR) approach reduces time for data capturing process, especially for big statistical activities. For National Statistical Offices (NSO's) there is always a trade between timeliness and data quality. Regarding this, an analysis has been made to estimate the data quality, for Enterprise Census.

When analyzing scanned data, there are several ways to divide the amount of data by questionnaire, page, field or character level. It was chosen the most detailed level, character level, where registered characters, in this case in the transfer file, is compared with characters from questionnaire. Boxes were picked randomly and from these boxes there were questionnaires randomly picket out. All characters were compared and looked for remaining substitutes, i.e. characters incorrect interpreted or incorrect registered.

Figure 5 shows how the interpreter interprets “6” as “0”. If the verifiers have accepted the “0” or change it to e.g. “8”, this was a remaining substitute. The fields were divided into three different kinds of fields; handwritten, checkbox and Optical Character Recognition (OCR) (i.e. pre-printed).

Figure 5: Example of a substitute



Status: Uninterpretable

450

Source: Enterprise Census, data capturing process

In Table 2 are shown the number of controlled questionnaire in order to analyze the scanned data for the Enterprise Census.

Table 2: Controlled questionnaire

Box Number	Number of questionnaires	Number of checked questionnaires	Share of checked questionnaires
4101	289	14	4.84
4601	252	7	2.78
5101	245	12	4.90
5501	296	8	2.70
6201	162	4	2.47
3902	259	14	5.41
4702	68	8	11.76
Total	**1571**	**67**	**4.26**

Source: Authors’ calculations

In the Table 3 are shown the number of substitute’s characters in order to analyze the quality of data capture for the Enterprise Census.

Table 3: Result of substitutes

Type of fields	No of characters	No of substitutes	Share of substitutes
OCR numeric	1474	0	-
Handwritten	18610	28	0.15
Checkbox	4110	1	0.02
Total	24194	29	0.12

Source: Authors’ calculations

The main purpose of controlling output data was to measure the quality by counting substitutes. Few numbers of substitutes were found as is shown in the Table 3. Handwritten fields represented most of found substitutes, which was not much of a surprise. A typical example was the letter "I" that had become a number "1" instead. OCR fields had no substitutes at all, only one occurrence of substitute for the checkboxes was found. The Table 3 reflects a very good quality of the captured/transferred data, only 0.12 percent of characters were substitutes.

2.5. QUALITY OF THE DATA

Users of survey data should always have at least some basic information about the degree to which survey data were modelled or estimated by imputation. At the end of imputation, it may be useful to produce the following indicators (Statistics Canada, 2010):

- the number of records which were imputed (i.e., the number of recipient records);
- the number of times each field was imputed and by what method;
- the number of records eligible to be used as donors;
- the number of records actually used as donors and the number of recipients each of these donor records imputed;
- a list (or file) indicating which donors were used for each recipient (to trace the sources of unusual imputed records);
- a list of all records for which imputation failed (e.g., because no donor was found).

Eurostat proposes the following standard quality indicators that can be used from the point of view of the producers, for summarizing the quality of the statistics as reported according to the Standard Quality Report (ESTAT/02/Quality/2005/9/Quality Indicators). This set of indicators can be used to measure and follow over time the quality of the data produced in the European Statistical System (ESS)[1]. The standard quality indicators proposed by Eurostat are:

- User satisfaction index
- Rate of available statistics
- Coefficient of variation
- Unit response rate (un-weighted/weighted)
- Item response rate (un-weighted/weighted)
- Imputation rate and ratio
- Over-coverage and misclassification rates
- Geographical under-coverage ratio
- Average size of revisions
- Punctuality of time schedule of effective publication
- Time lag between the end of reference period and the date of first result
- Time lag between the end of reference period and the date of the final results
- Number of publications disseminated and/ or sold
- Number of accesses to databases
- Rate of completeness of metadata information for released statistics
- Length of comparable time-series
- Number of comparable time-series
- Rate of differences in concepts and measurement from European norms
- Asymmetries for statistics mirror flows
- Rate of statistics that satisfies the requirements for the main secondary use

The set of indicators were considered by INSTAT to perform the assessment of the effects of the cleaning procedure (editing and imputation) at aggregate level, and can be grouped into three different kinds (INSTAT Quality Dimensions, 2014):

1. Indicators on the amount of data submitted to the imputation procedure, like Number of Records, Number of Variables, Number of Variables subject to the Imputation procedure, and Number of Total Values.
2. Indicators for the evaluation of the overall effects of the imputation procedure in percentage:
- Imputation rate (I): (Number of Imputed values/ Number of Total values)*100;
- Addition rate (Ia): (Number of Additions/Number of Total values)*100;
- Modification rate (Im): (Number of Modification/ Number of Total values)*100;
- Elimination rate (Ie): (Number of Eliminations/ Number of Total values)*100.
3. Synthetic indicators on the imputation rate by records, like for instance Number of Records with Imputation rate greater than 2% and Number of Records with Imputation rate greater than 5%.

From the indicators shown in Table 4, the LFS has around 0.01% Imputation rate, so CAPI data collection method is a very good method to monitor and control the quality of the data. PAPI, as expected, is the method where imputation and editing have the highest figures, around 4.71% of the total values.

[1] https://circabc.europa.eu/webdav/CircaBC/ESTAT/prodcom/Library/25-Quality/quality_reports/reference_documents/STANDARD_QUALITY_INDICATORS.pdf

Table 4: Quality assessment indicators at aggregate level

	LFS CAPI	STS PAPI	STS Online	ICT Tablet
Number of Records	12,490	6,803	60	1,473
Number of Variables	164	59	59	93
Number of Total values	2,048,360	401,377	3,540	136,989
Number of Imputed values	**20,484**	**1,890,486**	**3,929**	**56,165**
Number of Additions	-	545,873	2,443	5,480
Number of Eliminations	-	8,028	-	16,439
Number of Modification	20,484	1,336,585	1,487	34,247
% Imputation rate (I)	**0.01**	**4.71**	**1.11**	**0.41**
% Additions rate (Ia)	-	1.36	0.69	0.04
% Elimination rate (Ie)	-	0.02	-	0.12
% Modification rate (Im)	0.01	3.33	0.42	0.25
% Non-Imputation rate	**99.99**	**95.29**	**98.89**	**99.59**
% of records with I greater than 2%	**5**	**115**	**26**	**112**
% of records with I greater than 5%	**5**	**13**	**18**	**9**

Source: Authors' calculations

It is obvious from the above data that the CAPI method benefits are not only on the processing time (as data entry is done during data collection process) but also on the data quality.

3. OTHER STATISTICAL PROCESS, BESIDES DATA COLLECTION

INSTAT is aiming at replacing SAS and SPSS software in favour of R statistical software package which is an open source tool capable of doing most of the same thing as the others. This shift is mainly due to the very expensive license costs. In the meantime, before the process is fully accomplished, there is still a need for use of SAS and SPSS. So the plan is to gradually switch from SAS/SPSS to R.

In September 2012 INSTAT has developed a new well-functioning website with user friendly structure. It means using the Common Nordic Meta Model (CNMM) to allow user's access to statistical data, dynamically. INSTAT's statistical databases have been built with a PX-Web user interface. From January 2016 this tool is free of charge, while PX-Web is free of charge since January 2015. PX-Web is developed and owned by Statistics Sweden and is used to establish dynamic tables[2].

Every previously mentioned effort is done to contribute to the development of a sustainable statistical system in Albania. This system should facilitate decision-making based on relevant and reliable statistical information that meets domestic needs.

[2] http://www.scb.se/sv_/PC-Axis/Programs/PX-Web/

4. CONCLUSIONS

PAPI method has a good response rate; the interviewer has a positive effect on this aspect. INSTAT interviewers are equipped with additional materials (maps, formal papers, etc.) in order to improve the quality of data. Based on the quality assessment, complex questionnaires (skipping rules, rosters), are difficult to be completed.
CAPI method benefits from the positive effect of using interviewers, especially in response rates. INSTAT interviewers are equipped with additional materials (maps, formal papers, etc.). The data entry process is done during the data collection phase and more complex questionnaires can be used. Regarding the quality of the collected data, this method reduces interviewer's errors, as range and consistency rules are applied during the data collection. Equipment used for this method, even though prices are declining, is still costly. Using laptops or other kind of equipment in fieldwork is not very easy and a good training is needed. Interviewers and respondents have reacted positively to the use of laptops and tablets.

The data produced for this analysis support the assumption that the cost of laptops or tablets used for data collection process is divided by different surveys that use this equipment. The online method has the lowest cost and this method has also the lowest response rate. The burden on interviewers is higher on the online method, due to the fact that it requires more time to read and understand the questions.

CAPI and online method allow instant data transmission to headquarters, making possible immediate further processing of data.
Including innovation in the process of collecting data for the production of official statistics should be done by analyzing cost, quality of the data and processing time.

BIBLIOGRAPHY

Bradburn, N. M. "Respondent Burden", Paper presented at the 138th Annual Meetings of the American Statistical Association, San Diego, CA., 1978

ESTAT/02/Quality/2005/9/Quality Indicators,

Fannie Cobben, Statistics Netherlands, Nonresponse in Sample Surveys, 2009

https://circabc.europa.eu/webdav/CircaBC/ESTAT/prodcom/Library/25-Quality/quality_reports/reference_documents/STANDARD_QUALITY_INDICATORS.pdf

INSTAT, Quality Dimensions of the 2011 Population and Housing Census of Albania, May 2014

Statistics Canada (2010) Survey Methods and Practices. Statistics Canada, Catalogue no. 12-587-X, Ottawa. http://www.statcan.gc.ca/pub/12-587-x/12-587-x2003001-eng.htm