UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Working paper
13 February 2015

Workshop on the Modernisation of Statistical Production
Meeting, 15-17 April 2015

Topic (ii): Enterprise Architecture and its role in the Modernisation of Statistical Production

HARZEMLI, THE DDI BASED STATISTICAL PRODUCTION PLATFORM

Akın ÖZTÜRK, M.Sc.; Murat ÖZYÜREK, M.Sc.; Murat TUNÇEL, M.Sc.; İlker GÜVEN ,M.Sc.

akin.ozturk@tuik.gov.tr, murat.ozyurek@tuik.gov.tr, murat.tuncel@tuik.gov.tr, ilker.guven@tuik.gov.tr

Turkish Statistical Institute, IT Department

Abstract

This paper presents the recently developed, stunning statistics production platform, namely Harzemli, of Turkish Statistical Institute (TurkStat) that has been designed to generate a standardized approach to statistics production with an increased level of quality and decreased workload. This paper is within the context of Enterprise Architecture and its role in the Modernization of Statistical Production.

Harzemli is an application platform that mainly depends on DDI and rule XML files. The main concept behind Harzemli Platform is the standardization and modernization of the statistics production process through the usage of DDI files and rule XMLs in generating generic pages for the surveys, analyzing the collected data and dissemination of the data. Harzemli Platform automatically creates data entry applications from metadata. This platform has been created to standardize all statistical IT processes with respect to metadata. Metadata is actively used in production, analysis and data mining and dissemination phases of GSBPM model. Currently, in TurkStat, 83 out of a total of 92 surveys are being performed on this platform.

I. Introduction

1. TurkStat is the producer and coordinator of official statistics in Turkey with a precise mission to present high quality statistics. Harzemli Platform is the new DDI-based statistical production platform of TurkStat that provides a generic application development environment. Receiving as input a DDI file that contains reference and structural metadata information about the related study, it generates the respective survey application which actually has a generic user interface. This eventually results in the desired condition that the dynamically created webpages for different surveys have very similar look-and-feels.

2. Harzemli platform has been designed for the standardization and modernization of the statistics production process. It increases the reliability and consistency of data for official statistics production. It is a platform that minimizes the time to produce data effectively.

3. With respect to the dissemination phase of GSBPM, the end users are able to make dynamic queries on TurkStat's website in addition to the static tables and bulletins. Turkstat’s new dissemination system, namely MEDAS, consists of a common database and a common application and is sufficient for querying all different statistical subjects at the same time. This unique feature can lead the users to find new correlations between subjects.

4. Harzemli Platform consists of Metadata Editor, Rule Editor, Harzemli Desktop, Harzemli Web, Harzemli Mobile, Harzemli Management, Harzemli Analysis, Harzemli Data Visualization and MEDAS projects.

5. **Metadata Editor:** Metadata Editor is a tool that is used to create metadata in DDI format. Surveys can be designed on this software by defining the structural and referential metadata. Structural metadata consists of datasets, variables and variable groups. Document description and study description constitute the referential metadata.

6. The output XML file of Metadata Editor is the DDI file that is used as input by Harzemli Web, Harzemli Desktop and Harzemli Mobile applications to automatically generate the survey pages. Using this XML file, web pages that are used for data collection for different surveys eventually have similar layouts and look-and-feels.

7. **Rule Editor:** Rule editor is a desktop application that is used for defining the rules for a variable or rules that takes place between the variables. Rule editor is used to create simple edit rules, show-hide rules, functional rules, read only field rules and nullable fields rules.

8. The output XML file of Rule Editor is used as an input to Harzemli Web, Harzemli Desktop, Harzemli Mobile and Harzemli Analysis applications. Using this output XML file throughout data collection phase ensures that erroneous data is not allowed inside production database and even if erroneous data somehow gets inside, Harzemli Analysis application eventually finds it using this file.

9. **Harzemli Desktop:** Harzemli Desktop is a desktop application that is designed to work offline (without any internet need) to collect data using netbooks/notebooks. It is eventually the first data collection project designed in Harzemli Platform. It is designed to transform the medium of data collection from paper based to electronic platform. It receives the DDI and rule XML files as input and generates the separate data collection user interfaces for separate survey studies. The main benefit of Harzemli Desktop is that it allows to collect data on electronic platform without requiring constant internet connection.

10. **Harzemli Web:** Harzemli Web is a web application that transforms data collection survey forms which are prepared in DDI format into respective data-collection applications.

11. Through guidance of MetaData and Information Technologies Departments, the production units(departments) are fairly involved in the design of the layout and appearance of their respective surveys. Harzemli Web dynamically generates the survey web pages at run-time through processing DDI and Rule XML files, so production units can decide the design of the webpages through designing these files. Units can decide how many pages the

survey should take, how many parts it will involve, the length of textboxes, usage of radio buttons and so on.

12. **Harzemli Mobile:** Harzemli Mobile is an Android application that is designed for being used on tablets. The aim of Harzemli Mobile is to take advantage of mobile operating systems and leightweight devices on the field. Similar to Harzemli Web, through processing DDI and rule.XML files as input files, Harzemli Mobile generates data collection applications dynamically. Harzemli Mobile is able to save data offline without any need for internet connection. Later on, as the internet reconnects, it is possible to transfer the data on the tablet to central database through web services.

13. **Harzemli Management Console:** Harzemli Management Console is a web application that is used for authentication, authorisation, notification through SMS, reporting and planning purposes related to surveys that are handled by Harzemli Platform. Through processing the related DDI and rule XML files of the surveys as input files, Harzemli Management Console automatically generates the necessary database tables. It is also used for managing sample records. Any special purpose need with respect to surveys is covered by Harzemli Management.

14. **Harzemli Analysis:** Harzemli Analysis is a data analysis tool that allows the users to perform different ways of analysis on data through different technologies. Users of this analysis tool are able to create and run their own error-finder rules (through automatic transformation from logical expressions to SQL statements), run special rules (through automatic transformation from XML based rules to SQL statements) and trigger and run analysis on streams/files from other statistical analysis systems (including SPSS, SAS and R) for advanced analytics.

15. **Harzemli Data Visualization:** Harzemli Data Visualization is an application that allows statistical analyses with the graphics generated upon the data which was gathered with Harzemli Platform. With this application, SQL sentence which is generated according to the table selected by the user and column values of this table are sent to R server. Visualization of data is implemented according to selected data type and parameters sent by user and presented to user.

Data Dissemination (MEDAS):

16. It is known that the number of printed copies is diminishing day by day. The fact that Statistical Yearbook prepared by Statistics Canada will no longer be published is a strong example for this. Electronic dissemination may probably be the only main channel in the near future. There are various electronic dissemination channels in a typical statistical institute; press releases, predefined tables, dissemination databases (databanks), microdata files, thematic publications, etc. Among these, dissemination databases in Turkstat can be defined as aggregated data repository that is populated from microdata warehouses. They are mainly considered to have been developed for researchers, miners, or generally classified as intermediate or high level users. Known advantages of databases can be classified as:

- Avoiding duplication of information
- Helping to develop web applications
- Security
- Quick access to updated data

- Multilanguage
- Enforcing the usage of metadata

17. Turkstat used to have dissemination database for more than a decade. However, the strategy for adding a new subject in the dissemination database -from the ICT Department perspective- included assigning a database specialist to the subject. He/she starts with some meetings with the subject matter unit. He/she prepares his/her own tables, totally specific to the subject. There is no relation with other subjects' codes or there is no central metadata repository that could lead the end user to the correlated subjects. Hence, the end user must work hard to find whether there is a correlation or not, since he will see different reports and try to merge them manually in a single spreadsheet.

18. One other handicap of the old system is the web application part. As mentioned above, each subject has its own database tables in different structures. This not only brought maintenance problems to the ICT department, but also resulted in different Java application for each subject, which is again a factor that adds maintenance problems and person-dependency. End users had to deal with various web applications although there was a hard work in ICT part to make them look as if there is a single application. However, there was a lack of consistency between applications that can be understood when querying different databases.

19. Moreover, opening the new data to the end users involved a manual process. The subject matter unit staff calls or e-mails the database specialist saying that the related press release is released and data can be shared with end users. After that, the database specialist loads the final data to the dissemination database. It can easily be deduced that there were different data flows, wikis, codes for each subject.

20. One of the main reasons for the change was the technology used for dissemination databases was old and vulnerable to security threats, as well.

21. ICT Department conducted a study with the mottos of standardizing all the above-mentioned processes and not to involve in the content. The database design was the challenging part, since all datawarehouses start with a specific subject. We wanted to design a new approach that could be generic and that could let someone to see a single report regardless of the number of subjects chosen. We also wanted this model to let the data be served in a modern pivottable tables.

22. The role of metadata for internal processes is comprehended in Turkstat, especially since 2012. The writers of this paper believe that this comprehension is merely due to the traineeships to the EU countries. Therefore, we also wanted this design to be metadata based so that it could carry us to the next decade. Turkstat has a classification server for a decade. We searched for its data structure and decided that the metadata of MEDAS, the new data dissemination system, could be based on our existing classification server.

23. Some subjects were chosen as pilot by the presidency. The results of these pilot projects were quite satisfactory both by the statisticians and the board. MEDAS is in use since April, 2014. Hence, Turkstat is now migrating its old fashioned dissemination databases to MEDAS. All new subjects that will be served via databank are also developed by MEDAS. The data structure of MEDAS has been taught to the related staff of all subject matter units

via workshops. To sum up, regarding the background processes for a databank, with MEDAS:

- ICT department just builds the pipe line and the data is filled with subject matter units. This reduces the burden on both ICT staff and ICT directors.
- There is only one dissemination schema and one database design that includes all the dissemination data.
- There is only one databank application. However, some other practical applications are and will be written based on MEDAS data. MEDAS reports reduce the reporting burden on ICT staff.
- Users are now able to use modern pivot reports and compare any number of subjects in the same report page. The latter is believed to be an innovation in statistical IT world.
- Manual intervention of waiting for a call or e-mail from the subject matter unit saying that the dissemination could be uploaded is history. MEDAS waits for the related press release (if any) to be released and the data is shown to the end users afterwards.
- Uniform data model makes it easy to develop web service.

24. We believe that MEDAS, especially with its data design, is an innovation for statistical production processes and some developments are being conducted for serving the data with graphs and thematic maps.

II. Methodology

Harzemli Project Objectives

25. The main objectives of Harzemli Platform are

- Shortening the process of coding the data entry program,
- Enabling data entry program to not being dependent on the developers
- Developing data entry program with standard application codes.
- To standardize the names of all the variables in terms of data integrity
- Ensuring faster compilation of private sector data entry by respondents in the survey and the immediate correction of erroneous data input

26. The model may be a reference to other public institutions and all survey / data collection process of public organizations is aimed to perform in accordance with international standards

Working Methodology and Techniques

27. **Feasibility study:** By analysing an XML based generic application development platform, named “wizard”, it is decided that Harzemli Project, an advanced version of this “wizard” application, can be developed with current technological infastructure being used in TurkStat.

28. **Cost benefit analysis:** Regarding the benefits of the possible outcome of the project, a benefit-cost analysis has been conducted. It is decided that instead of outsourcing or

purchasing any software, it is more feasible to create our own software with the efforts of staff at ICT Department.

29. **Risk Management Plan:** Risk prevention plan has been developed based on risk conditions which are established by using expert recommendations and experiences earned during the development of wizard application.

30. **Quality Management Plan:** Quality management plan has been defined with the aim of using DDI xml format in an efficient way and getting recent development on DDI standard.

31. **Change Management Plan:** Change management plan has been developed based on job steps which are required to be made and followed when a revision request made about survey studies.

32. **Human Resources Management Plan:** Human resources management plan that consists of defining project roles, assigning responsibilities and designing project organization schema has been developed. Number of Project team members has increased by using the context of human resource management plan after implementing first few modules of the Project successfully.

Project Milestones:

33. The project milestones are as follows:

	Milestones of the project	Development dates
1	Harzemli Desktop	1
2	Harzemli Rule Editor	2
3	Harzemli Management Console	3
4	Harzemli Web	4
5	Harzemli Mobile	5
6	Harzemli Analysis	6

Disclosure and analysis of requirements:

34. At the beginning of the project, Household Labour Force survey was selected as a pilot study since it is conducted in six month periods periodically. In order to understand application requirements in details, a joint working group has been built by both software development team members, statistical unit staff and users of the application with aim of defining the requirement analysis processes in details.

35. System and user requirements (functional, non-functional and quality requirements) are defined based on the meeting decisions regarding to business logic and data structure of the application. These meetings are organized by analysis team in a way that requests from the end users are added into the general requirements.

Project Management of Harzemli Platform:

- ✓ After completion of planning and analysis processes, project team members are assigned and process of development of the project is started.

- ✓ In management of the project, Agile Project Management method is used in a way that Project modules can be extended based on the user requirements. At first, Harzemli Desktop module is developed. By using experiences learnt from development of Harzemli Desktop, Harzemli Web and Harzemli Mobile modules are designed.

- ✓ In management of project requirements, the most complex and error prone requirements and components are selected to be developed because of encountering these possible errors in the beginning of the development of the project. Since household labor force survey contains most of the requirements of the other studies and it has a complex workflow, the first study which is decided to be implemented by Harzemli Project became household labor force study.

- ✓ Defining changing requests and implementation of approved ones: Errors that are emerged as a result of not satisfying the requirements are detected during the quality control testing phase.

- ✓ In risk management, in order to prevent collecting incorrect or missing data, paper forms are selected as data collection method in first a few studies.

- ✓ The most complex studies are implemented by using Harzemli Project in developing Harzemli modules.

System Test

36. Testing activities in IT department is managed by defining test scenarios and tracking these test cases. Survey on Information and Communications Technologies (ICT) Usage in Household which is the first study implemented by Harzemli Desktop is tested by staff responsible with defining survey requirements in statistical unit, software test professional in ITC department and pollster who are responsible for collecting data via application. After completion of testing phase, most of the bugs and inaccurate workflow of the application are fixed by project team members before application is used in real time environment. Each new release of application is tested by both testing staff and statistical unit staff. In order to run a performance test on the application which is used by lots of users, it is tested by simulating users generated by test simulation tools.

III. Results

37. In 2013, 32 survey studies were performed through Harzemli Project. Out of these 32 studies, 6 were performed through Harzemli Desktop and 26 were performed through Harzemli Web. As of 2014, the total number of survey studies through Harzemli was 82. For the year 2015, 22 more survey studies are planned to be included in Harzemli and studies are ongoing for this aim.

	Harzemli Desktop	Harzemli Web	Harzemli Mobile	
2013	6	26	*	32

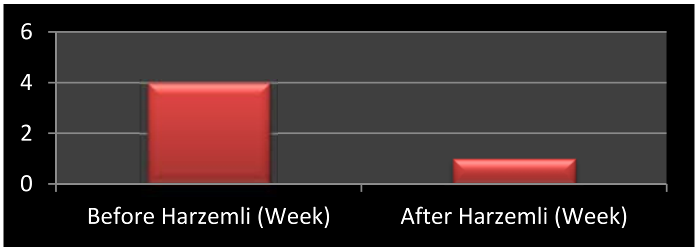
2014	4	46	*	50
2015	2	13	7	22
TOTAL	12	85	7	104

Productivity Gains

1. Time

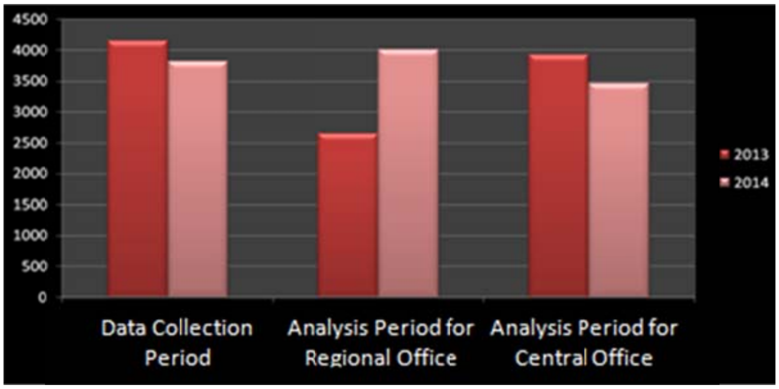
a. Decreased time for software engineers to develop data entry applications

For every survey study conducted by the TURKSTAT; study-specific software development requirement has been eliminated. Depending on the type of the study, data entry program for any study is now automatically created through one of the three data entry applications (Harzemli Desktop, Web or Mobile applications) according to the XML files used for defining metadata and rules (i.e. DDI.xml and Rule.XML, respectively). While the average time necessary for developing a data entry program for a study with an average level of difficulty was 4 weeks, it is now only a week.



b. Decreased time for Data collection

Before Harzemli platform, surveys were pressed and transmitted to the respondents to collect data. Data entry was performed in regional offices through processing forms that had previously been filled in the paper medium by respondents. Using Harzemli Platform, processes have been coupled through collecting data directly from the respondents, thus saving time and labor. Data collection time has been shortened by 8% and thanks to instant access to data, analysis time available for regional offices has been increased by 50%. As the data comes directly from respondents to TURKSTAT and the officers from regional offices are now able to spend much more time on analysis, there has been a %12 drop in the analysis time that is necessary for the officers of central office.



c. Decreased time for preparation of press releases

Having the collected data transmitted instantly to central office and diminishing the necessary time for analysis, data dissemination has been quickened. The time period necessary to prepare the press releases has been shortened. For example, the time period that is necessary for the preparation of the monthly Labor Statistics press releases has been shortened by 4 days.

2. Quality

a. Software Process Standardization
With Harzemli project, standard software business processes have been established and job descriptions have been determined. Thus, control, consistency and order of processes are established and complexity is reduced. Dependency of software process on individual persons has been removed.

b. Data integrity
In order to generate data with Harzemli Project, usage of standard code lists and standard variable definitions ensures the integrity of data. Code lists and variables were not common for usage for each study before the Harzemli project. Through standardization, data communication between studies has been obtained. Thus, it was also made possible the development of data exchange and data distribution software at TURKSTAT.

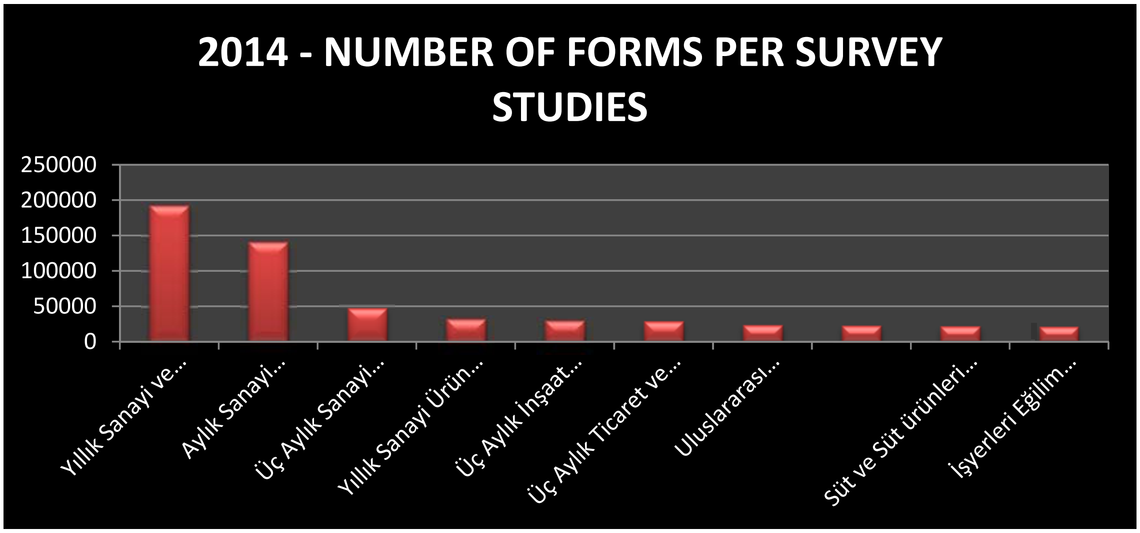
c. Common code development
Before Harzemli Project, applications for each study were developed using different programming languages and technologies. Therefore it was almost impossible to inspect the quality of the all the applications. Prior to Harzemli project, maintenance, quality assurance and software testing were to be done for 90 separate data entry applications whereas now this has been reduced to only 3 applications (namely, Harzemli desktop, Harzemli web and Harzemli mobile). Switching from individual code development logic to development logic as a team, more functional, reusable, multiple purpose applications with standardized interface have been developed.

d. Common Components Available
Codes written prior to Harzemli project has been splitted into modules as the project has progressed and layering has been implemented. The software components have been optimized to be suitable for being commonly used in Harzemli Desktop, Web and Mobile applications and thus costs of seperately writing and maintaining components have disappeared. Having metadata standardization done, standard reporting mechanism has been accomplished under Harzemli Management Module and thus, the need to write seperate programs for the reports of follow-up and current status analyses for each application is eliminated.

3. Costs

a. Paper
In order to keep pace with constantly evolving technology in the world and in our country, reduce labor and time costs to a minimum level and produce data

in a timely manner; TURKSTAT has started to carry out all corporate based surveys through Harzemli Web application. The paper editions of surveys have been canceled. In 2014, thanks to Harzemli Project, 7,565,680 sheets of paper were saved from getting pressed for conducting all the surveys with a total of 756.568 respondents.



38. In 2015, the sample volume has reached the level of 2,045,974 respondents through Harzemli Web and Harzemli Mobile applications. Considering that an average of 10 pages are needed for conducting a survey with a respondent, 20,459,740 sheets of paper savings are anticipated.

IV. Conclusion

39. Harzemli Platform is an innovative DDI-based statistical production platform that has recently been developed in Turkish Statistical Institute and is currently being used in generating official statistics of Turkey. Before Harzemli Platform, staff at ICT Department had been assigned the implementation of seperate web applications for each survey. Due to the fact that engineers at ICT Department are generally from different backgrounds and have their own way of logic on software, the applications had different look-and-feels and there was no standard coding style. Harzemli Platform has provided a high level of standardization on data collection applications, thus, generating low-cost, better-quality and long lasting applications.

40. Harzemli Platform has been designed for the standardization and modernization of the statistics production process. It increases the reliability and consistency of data for official statistics production. It is a platform that minimizes the time to produce data effectively. Standardizing the names of all variables in terms of data integrity, software development with standard application codes, being not dependent on developers and shortening the software development process are among great profitable outcomes of Harzemli Platform. Using common classification and code lists, Harzemli platform standardizes data collection processes. It shortens the duration of preparation time for data entry programs from 40 days to 10 days, and shortens press release date for a minimum of 4 days.

41. So far, Harzemli Platform has had great success and has seen great support by both the presidency and staff from the units. It is anticipated that Harzemli Platform is going to enlarge by the new modules that are currently being developed to be added in the near future.