Distr.
GENERAL

WP.22
30 March 2010

ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Meeting on the Management of Statistical Information Systems (MSIS 2010)
(Daejeon, Republic of Korea, 26-29 April 2010)

Topic (iv): Innovation and related issues including census systems

Software Systems for Surveys and Censuses

Prepared by Yudi Agusta, Statistics Indonesia, Indonesia

I. Introduction

1. The conduct of survey or census, from a certain point of view, can be considered as a temporary matter that can be conducted in a temporary manner and finished roughly as the time of survey or census ended. Extensive changes/consolidation in terms of concept and definition, scope, and uses of variables that need to be taken during the time of survey or census conducted are the reasons for this. Some problems are usually involved in accommodating the changes/consolidation needed. Miscommunication between local and central administrators, misinterpretation, and communication lag are among the problems arising from this matter.

2. Among other problems, which are related to one another, documentation in regards to the changes/decision taken during the survey or census has also contributed significantly to the results of data analysis conducted after the survey or census finished. As the survey or census getting busier, administrators often forget to record the changes/consolidation they have made. Bad documentation will eventually result in a bad analysis since the interpretation of recorded data is not properly or appropriately delivered.

3. Some other problems related to the conduct of survey or census include lack of accuracy of the recorded data, handling problems of survey or census documents, and lack of the use of current technology in exploring and comparing the collected data. The first problem, which is the lack of accuracy, often arises during the process of data collection which can involve false information from respondent or bad interpretation by enumerators. Such problem is often approached by strictly conducting a cross-examination process. However, such process is still unable to significantly improve the quality of collected data.

4. The handling of survey or census documents has also faced problems. Documents are often processed without documenting properly who processes the documents or which part of documents is currently being processed. The prepared manual handler has some drawbacks in terms of its convenience and accuracy. The use of current technology in exploring as well as improving the quality of the recorded data is often not optimal.

The current technology such as data warehouse systems[1,3,4], data analysis/data mining/intelligent systems[6,7,11], and online dashboard systems[5] are among new technology that can be utilized for those purposes.

5. Based on the problems and background mentioned above, a comprehensive software system for smoothly conducting surveys and censuses and improving the quality of collected data is highly needed. For that reason, this paper tries to elaborate a comprehensive analysis of the software systems for surveys and censuses. The paper is structured as follows: Section II elaborates on Base Theories and Background considered for the design, Section III discusses Analyses of Required Software Systems, Section IV and V explain respectively, some Hurdle faced for the implementation of the design and Further Advancement of the design, and Section VI provides a Conclusion of the paper.

II. Base Theories and Background

A. Information Systems

6. Computer science so far has resulted in an enormous valuable theories and knowledge that can be used to support all aspects of human activities. Information system[5], in this case, is a system which has a task to connect computer science field and human as users of the knowledge included in the information system. Based on its use, there are a number of types of information systems commonly implemented including management information system, decision support system, executive information system, data warehouse system, intelligent system, and geographical information system, among others. All of them are generally useful in supporting the human daily activities.

B. Intelligent Systems

7. Intelligent system[6,7,9,11] is a system which applies a human-like intelligence so that it can solve problems intelligently within a specific domain included in the system. Among them, there is Rule-Based Systems which uses systematically recorded rules for solving problems. The rules involved in a rule-based system can be in the form of a decision support system[9] where the rules are extracted from an expert and engineered into the form of recorded rules or an engine which is based on some intelligent agents[6,7,11] such as clustering agent, natural language processing agent, association rule agent, and pattern matching agent, among others. Intelligent system has extensively used the research results obtained from Data Mining[6,7,11] and Soft Computing fields.

8. Among concepts available, decision support system[9] is one type of intelligent system that has been widely used. Decision support system provides user with capabilities to trace rules, relation between rules, and eventually an alternative answer of the occurring problems, so that user can make a decision based on the advises provided by the system. Rules included in the system can be obtained from expert knowledge or from the data available for the problem.

C. Data Warehouse Systems

9. Data warehouse systems[1,3,4] is a system for storing aggregated/summarized data which can be analysed online from various points of dimensions using an Online Analytical Processing (OLAP) engine utilizing slice and dice or roll up and drill down facilities. Aggregated/summarized data as well as field dimensions involved in a data warehouse design have to be extracted, transformed, and loaded from their source systems, which are usually in the form of Online Transaction Processing (OLTP) system, into the database of the data warehouse system, using an Extraction Transformation Loading (ETL) tool prepared specifically for the design.

D. Web-Based Online Systems

10. Web-based online system[2,10] is a system that can connect one system to another in real time and utilizes web technology. The system is often represented by the use of hyperlink that can move user from one

interface to another. Web-based online system has developed fast, been used widely and is very useful to effectively spread information to a large number of people.

E. OCR and OMR Technology

11. Optical Character Recognition (OCR) Technology[8] is a technology used to recognise character in the form of images. The technology has developed fast and shows capabilities in recognizing character images accurately, even if the images are full of noises. Methods applied to the technology has been varied including image processing, pattern recognition, soft computing method, noise reduction method, and rule-base method, among others.

12. Optical Mark Recognition (OMR) Technology[8] is a similar technology to the OCR technology and has been used for long time to process documents with mark images as inputs. The recognition accuracy of this technology has been very good. For that reason, the technology is highly recommended used for documents processing.

F. Currently Implemented Software Systems

13. For supporting the conduct of survey or census, Statistics Indonesia has developed a number of separated software systems for processing questionnaires collected from the field and monitoring the progress of field enumeration as well as data processing. These software systems have been developed following the requirement and are often based on a temporary analysis, design, and implementation. A settle and comprehensive approach for comprehensively handling the problems involved within the conduct of survey or census yet to be developed or implemented.

14. The technology used for the development includes client-server technology, web-based online systems, and the need basis data mart systems, among others. The technology used is usually chosen following the basic needs, the availability of in-house programmers' expertise, and not based on a comprehensive analysis. This lack of comprehensive analysis causes a lot of problems in applying the systems for survey or census, since the systems developed is often not user friendly, prone to error, slow in running, or lack of the needed utilities.

III. Analysis of Required Software Systems

15. The analysis of software systems reported in this paper contains an analysis of a ready-to-use software system for any type of survey or census. Adjustment is only needed for the difference in terms of the types and the number of variables included in the survey or census. The analysis will involve a number of modules including an Online Interaction Module, a Document Handling Module, a Data Processing Module, a Post Processing Utility Module, and a Data Mart-Based Data Exploring Module.

A. Online Interaction Module

16. A number of problems related to the interaction between local and central, field enumeration and data processing administrators include:

(a) Interaction between local and central field enumeration administrators in terms of concept and definition, scope, uses of collected variables is not properly conducted.
(b) Interaction between local and central data processing administrators in terms of error of systems, and utilities setup within the systems is not properly conducted.
(c) Interaction between local and central field enumeration and data processing administrators in terms of validation rules implemented in the developed data processing systems is not properly conducted.
(d) Progress reports both for field enumeration and data processing stages are not properly presented with analysis whether the progress is on track or not.
(e) Patches and other supporting files are often sent using informal methods which can cause the lost of track of the needed patches and files.

All the problems mentioned are generally caused by the unsettle way of communication between local and central administrators which eventually causes miscommunication, misinterpretation of communication, communication lag to happen. The method for broadcasting any announcement of changes/consolidation made also has an effect on these communication problems.

17. Figure 1. shows the Context Diagram of the proposed Online Interaction Module. Facilities set up in the module include:

(a) Field enumeration related question and answer facility for accommodating discussion and decision making process in regards to problems arising during the field enumeration stage.

(b) Data processing related question and answer facility for accommodating discussion and decision making process in regards to problems arising during the data processing stage.

(c) Validation rule suggestion and confirmation or approval facility for accommodating discussion, decision making, and inclusion of validation rule into the data processing module.

(d) Field enumeration progress reporting facility for uploading progress report and monitoring progress of field enumeration.

(e) Data processing progress reporting facility for uploading progress report and monitoring progress of data processing stage.

(f) Patches and other supporting files upload and download supporting facility for updating the facilities included in the software systems.

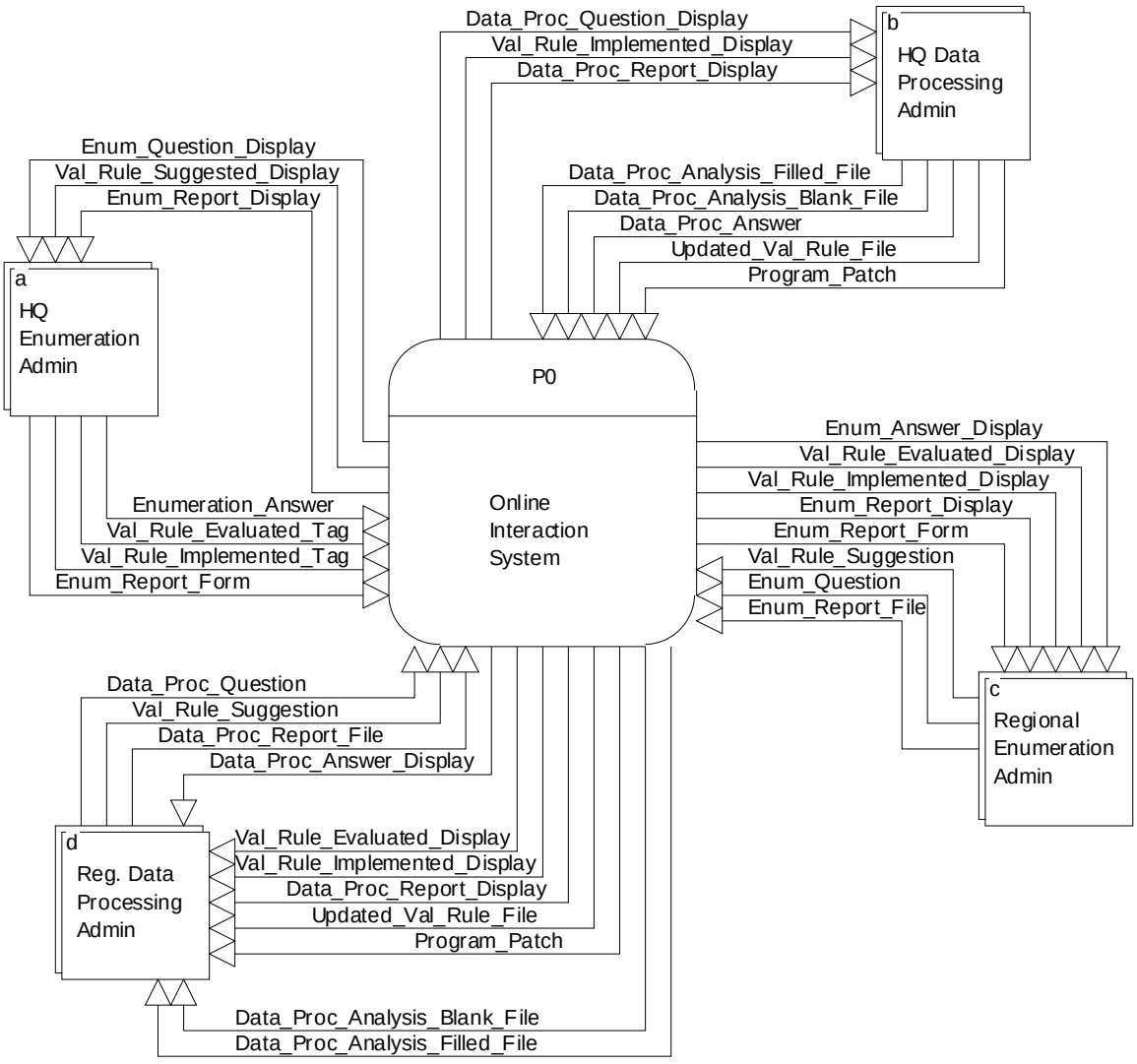

Figure 1. Context Diagram of the Proposed Online Interaction Module

18. Some extension notes on the module include:
   (a) The module is developed online applying the web technology.
   (b) The module involves interaction facilities for local and central field enumerator administrators only, local and central data processing administrators only, and between local and central field enumerator administrators and data processing administrators.
   (c) Decisions/changes made during the time of survey or census are effectively recorded and easily accessed by related parties.
   (d) The module can also be implemented in the form of online dashboard system, so that any occurring changes/consolidation taken can be monitored in real time.
   (e) Facilities such as analysis of progress and list of patches and other supporting files are included in the module.

B. Document Handling Module

19. Problems involved in document handling are bad and ineffective method used for record keeping of documents used in each data processing stage including receiving batching process, editing coding process, data entry process, cutting process, scanning process, correcting and completing process, and validating process, among others. Record keeping process so far has been conducted manually and mistakes in filling in the record keeping form can easily occur. For that purpose, an automated module for handling survey or census documents is needed to effectively keep the track of documents during the process.

20. Figure 2. shows the Context Diagram for the proposed Document Handling Module. Some notes on this design include the consideration of the document storage capacity, and the processes included in the data processing. For some cases, where the documents storage spaces are available separately in a number of different places, the system needs to be developed online. If not, the client server system will be enough. The processes included in the systems will depend on the data processing method applied, whether it is a conventional data entry system or a more advanced system such as an Optical Character Recognition (OCR)/Optical Mark Recognition (OMR) technology-based data processing system.

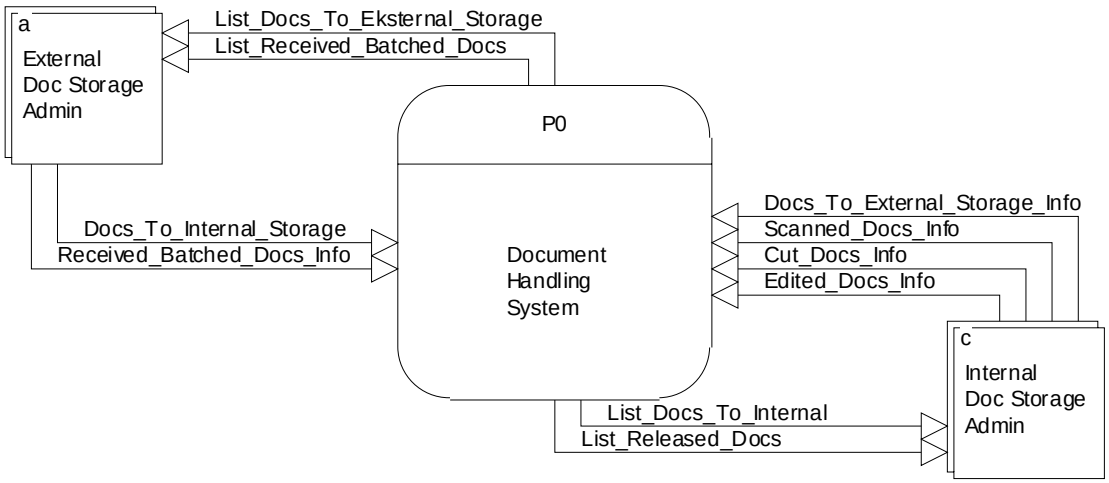

Figure 2. Context Diagram of the Proposed Document Handling Module

C. Data Processing Module

21. The data processing module is the core module of the software systems which aims to process questionnaires filled in during the field enumeration stage into the form of computerized digital data. The module will strictly follow the applied data processing method. In the case of Indonesian 2010 Population

Census, the technology implemented for the data processing is the OCR/OMR technology, so that the design of the module will be roughly as shown in Figure 3.

22. The module is designed in the form of a rule base system using the Validation Rules File as a rule source database for the system. A decision support system method which utilizes its knowledge base and inference engine components and has the ability to confirm relation between variables as well as find the unmatched data entered by operators can be the solution for that purpose.

23. For the OCR/OMR technology, since the storage and networking capacity needed for storing and transferring images during the data processing is large, the architecture of networking system used to support the process has to be arranged appropriately in advance. In this case, servers used for each process might need to be set up separately.

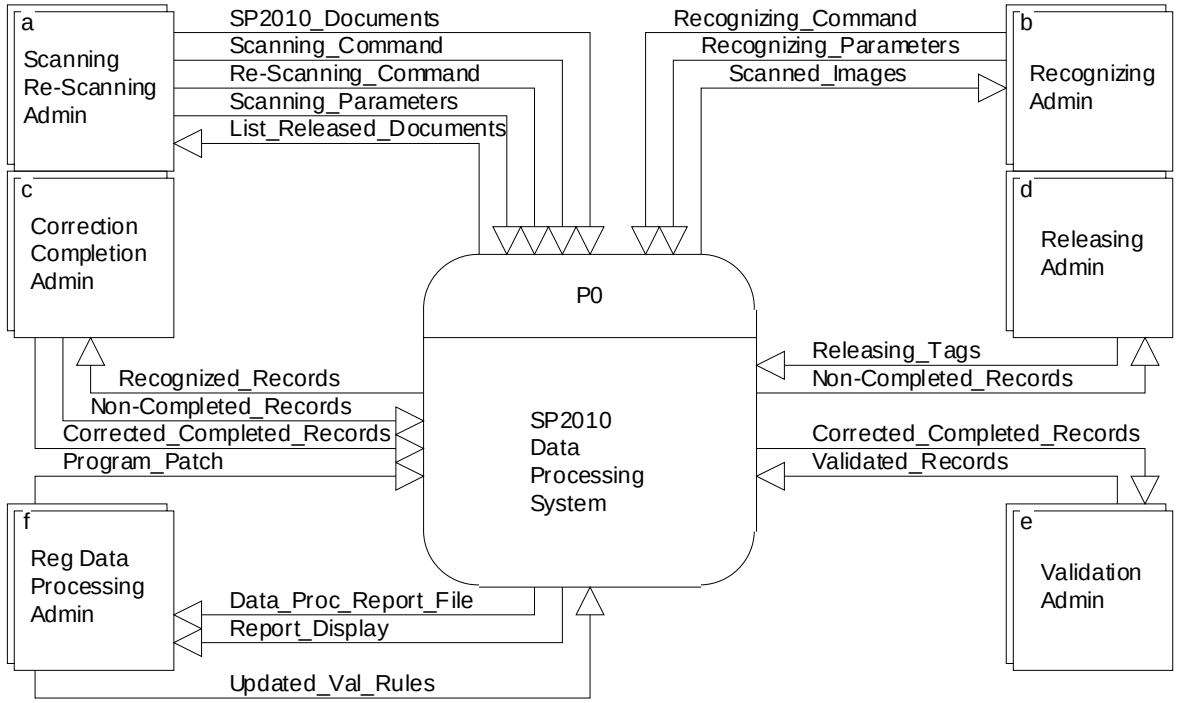

Figure 3. Context Diagram of the Implemented OCR/OMR Technology Based Data Processing Module

D. Post Processing Utility Module

24. Some types of variables included in a survey or census such as name, address, relation between name and sex, age in relation to other variables, record outliers, and record similarity, among others, cannot be validated easily and automatically by an application program. They usually need to be validated manually by an editor. This can cause a lot of delays in speeding up the data processing, since a number of experienced editors needs to be hired and time required for the editing process can be very long.

25. This post processing utility module is an additional advanced module for analyzing those recorded data that cannot be validated based on the validation rules implemented in the data processing module. The uses of some intelligent agents are very useful for that purpose including the natural language processing analysis, cluster analysis, and association rule analysis, among others. In the case of the up coming Indonesian 2010 Population Census, some types of intelligent analysis can be performed including name analysis, address analysis, sex analysis, age analysis, double counting analysis, and individual record outlier analysis, among others.

26. Figure 4. shows roughly the module designed for the post processing stage. Some notes on this design include that a specific expertise or knowledge in analysing, designing, and implementing the intelligent agent into the module is appropriately needed.

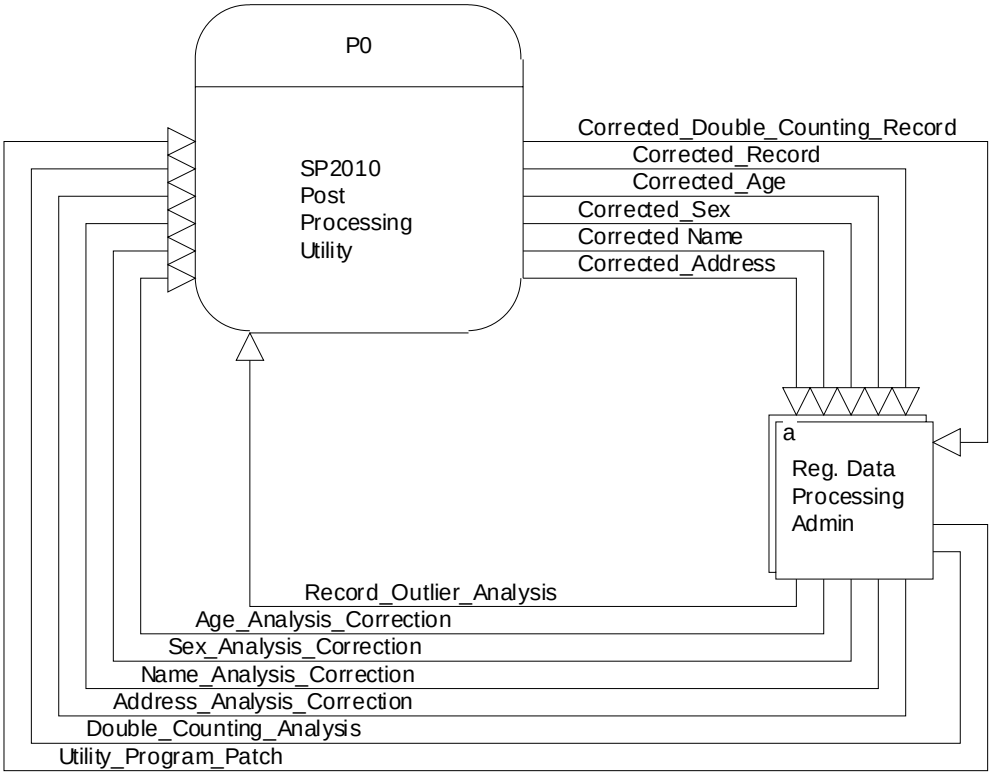

Figure 4. Context Diagram for the Proposed Post Processing Module

E. Data Exploring Module

27. Comparison of recorded data between times and area can be a solution for analyzing the normality of data recorded in each time or area. Data warehouse technology provides a solution in easily performing online analysis through OLAP (Online Analytical Processing) technology. This module can be both used for confirming the validity of the recorded data as well as for dissemination and advanced analysis purposes conducted after all stages of survey or census are finished. The most common dimensions included in a data warehouse system are time and area. This is also the case in the conduct of Indonesian 2010 Population Census, since the census involves area variable, and the comparison between time of census can also be included. Other dimensions that can also be included are, for example, age dimension (age range and individual age), sex dimension, and education dimension, among others.

28. Figure 5. shows the Context Diagram for the proposed Data Exploring Module that is designed based on the data warehouse technology which involves Extraction Transformation and Loading (ETL) facility for extracting, transforming, and loading survey or census individual record into the data warehouse system and Online Analytical Processing (OLAP) facility for conducting area analysis, time series analysis, or simple slice and dice or roll up and drill down analysis.

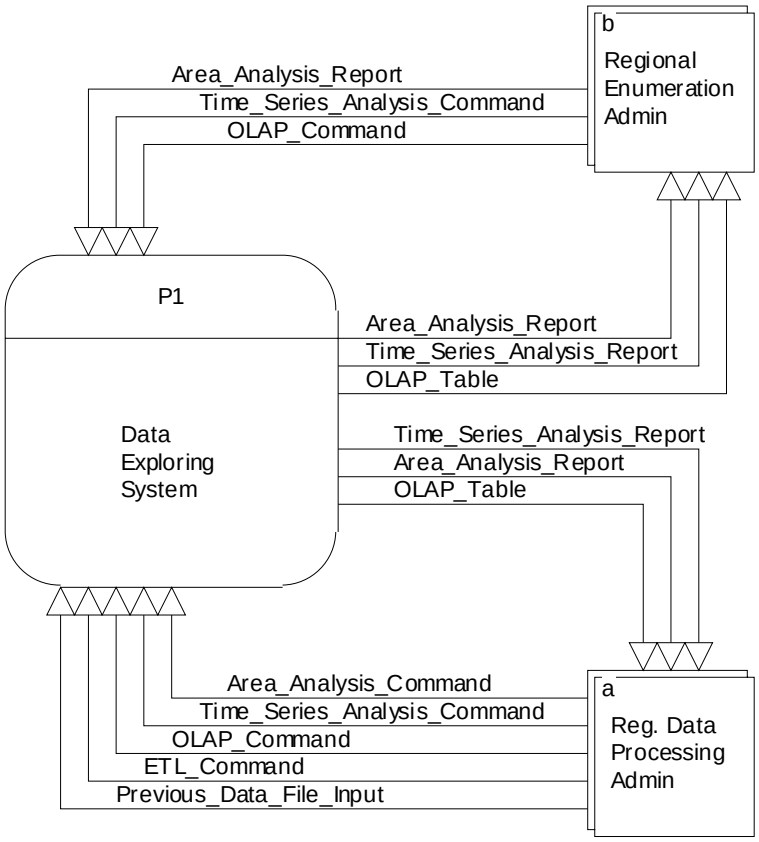

Figure 5. Context Diagram for the Proposed Data Exploring Module.

IV. Hurdle in Implementing the Software Systems

A. Information Technology Blue Print

29. In the mean time, there has not been an Information Technology Blue Print setup at Statistics Indonesia. This effect on the implementation of the software systems, since the software and hardware used for the implementation of the application can be different from that which will be used as standard for software development within the organisation. The software used will include a relational database management system, a programming application, web-development application, and stand alone or server operating systems, among others. The specification of computer and networking hardware can also affect how the software system will be implemented.

B. The Problem of In-House and Outsourcing

30. The design of software system presented in this paper is still partly up for implementation. Considering the schedule of activities setup within the organisation and the number of in-house programmers available, it can be considered that the scope of this project is too big to be handled by in-house programmers, if it needs to be implemented in a short time. Setting up in-house programmers dedicatedly for implementing a large design of software systems is also still an ongoing problem within the organisation. In this case, the software systems have to be implemented step by step and module by module following the availability of in-house programmers and the activity condition within the organisation. On the other hand, if funding is available, implementation of the design by outsourcing can also be the right solution for the problem.

V. Further Advancement

31. The software system can also be added with some facilities for supporting project managers' activities in the form of managerial support system. This can include an activity scheduling module for effectively arranging managerial schedule between managers, and a managerial decision support module for effectively making decision during the conduct of survey or census, among others.

32. The software system can also be extended into the form of an engine to develop a document handling module, a data processing module, a post processing utility module, and a data exploring module which can have a flexibility in accommodating the difference in terms of the types and the number of variables included in survey or census.

VI. Conclusion

33. Some conclusion can be taken from the paper including:

(a) Software system for comprehensively handling the occurring problems and procedures taken in a survey or census is highly needed

(b) The software system can involve a number of modules such as online interaction module, a document handling module, a data processing module, a post processing utility module, and a data mart-based data exploring module.

(c) Some advance technology such as rule base system, intelligent systems, and data warehousing systems can be applied to the software systems.

(d) Some hurdle needs to be considered into the implementation of the systems

(e) Further advancement can be added to further handling all aspects involved in the conduct of survey or census.

VII. References

[1] Imhoff, C. (2003). Mastering Data Warehouse Design: Relational and Dimensional Techniques. Wiley, ISBN-10: 0471324213.

[2] Inmon, W. H. (1993). ORACLE: Building High Performance Online Systems. Wiley, ISBN-10: 047156740X.

[3] Inmon, W. H. (2005). Building the Data Warehouse, 4th Edition. Wiley, ISBN-10: 0764599445.

[4] Kimball, R. (1996). The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses. John Wiley & Sons, ISBN-10: 0471153370.

[5] Laudon, K. C. (2003). Management Information Systems, 8th Edition. Prentice Hall, ISBN-10: 0131014986.

[6] Mitchell, T. M. (1997). Machine Learning, New Edition. McGraw-Hill Higher Education, ISBN-10: 0071154671.

[7] Negnevitsky, M. (2004). Artificial Intelligence: A Guide to Intelligent Systems, 2nd Edition. Addison Wesley, ISBN-10: 0321204662.

[8] Spencer, H. (1999). Automated Forms Processing: A Primer : How to Capture Paper Forms Electronically and Extract the Data Automatically. CMP Books, ISBN-10: 1578200490.

[9] Turban, E., Sharda, R., and Delen, D. (2010). Decision Support and Business Intelligence Systems, 9th Edition. Prentice Hall, ISBN-10: 013610729X.

[10] Walker, G., Janes J., and Tenopir C. (1999). Online Retrieval: A Dialogue of Theory and Practice, 2nd Sub Edition. Libraries Unlimited, ISBN-10: 1563086573.

[11] Witten, I. H. , and Frank, E.(2005). Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition (The Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann, ISBN-10: 0120884070.