Managing Data Acquisition, Cleansing and Transformation in an Agriculture Data Warehouse

Ahsan Abdullah (Corresponding author)

Department of Information Technology, King Abdulaziz University

Jeddah, Kingdom of Saudi Arabia

Fuad Bajaber

Department of Information Technology, King Abdulaziz University

Jeddah, Kingdom of Saudi Arabia

Received: November 19, 2015 Accepted: December 3, 2015 Published: February 17, 2016

doi:10.5296/jas.v4i1.8583 URL: http://dx.doi.org/10.5296/jas.v4i1.8583

Abstract

Pakistan is the world's fourth largest cotton producer (Anonymous, 2015). The country relies heavily on cotton yield to sustain and enhance its export and economic growth. Several state run organizations have been monitoring the cotton crop for decades through pest-scouting, agriculture and meteorological data-gathering processes. This non-digitized and non-standardized dirty data is of little use for strategic analysis and decision support. This paper is based on the data collection and cleansing issues of that cotton pest-scouting data consisting of approximately 15,000 sheets from 20 cotton-growing districts of Punjab province. Various real-life agriculture data management and data quality problems are discussed and explained in this paper using several real examples.

Keywords: Agriculture, Pest Scouting, Pesticides, Data warehouse, Data Quality, Data Cleansing

1. Introduction

For decades, different government departments have been monitoring dynamic agricultural conditions all over the Punjab province i.e. the breadbasket of Pakistan (Davidson et al. 2000). Subsequently, hundreds and thousands of digital and non-digital i.e. conventional files are created from hundreds of pest-scouting, yield surveys, meteorological data recordings and other similar undertakings. The multivariate data collected from different sources is heterogeneous and dirty, which is difficult to integrate. The lack of data standardization, cleansing and integration contributes to an under-utilization of this valuable and expensive asset. This results in a limited capability and utility of this data to provide decision support, analysis, research and development (Abdullah et al. 2006). Data quality in agriculture has different facets too, such as field sensor data which is typically noisy [ifti15]. For such data, the popular data cleansing methods based on prediction, moving averages or classification are not suitable.

ADSS (Agriculture Decision Support System) aims at using agro-met data for analysis, decision support, research, development, education and ultimately, to solve agro-related problems. Conventionally agricultural analysis and decision making process is based on expert opinion. ADSS employs in-house cutting-edge IT tools and techniques and offers decision makers an advantage over traditional way of analyzing data. To achieve the objectives, data inscribed on pest-scouting sheets has been collected, digitized, cleansed and used in the agriculture Data Warehouse. This wealth of agricultural data is then presented to, and analyzed by ADSS users for their strategic analysis and decision support. ADSS has the analysis capability of using decades old pest-scouting data, but in this paper we will discuss the data for six years i.e. from 2001 to 2006 covering approximately 15,000 sheets with 244,000 records.

This paper is divided into five sections. Section 2 gives the background information and definitions to familiarize readers with relevant terminologies used in this paper. Related work is in Section 3. Section 4 discusses the data extraction and transformation along with data cleansing. The paper concludes with conclusions in section-5.

2. Background

2.1 Data Warehouse and Decision Support Systems (DSS)

A decision is a choice among alternatives which are centered on estimates of alternative values. Decisions are based on qualitative or quantitative approaches or their combination. This may also require using past experience and knowledge of the current situation. Figure-1 shows the decision making process (Heinemann, 2010). Decision Support Systems (DSS) are meant to make the corporate historical data available for decision makers in the course of strategic planning and analysis. DSS relies on a Data Warehouse for keeping summarized records and making the data available as a single source of truth. Supporting a subsequent decision means people work alone or in a group and gather necessary intelligence, generate alternatives and finally make choice(s). Supporting the choice-making process involves supporting the approximation, the evaluation and/or the assessment of alternatives (Daniel, 2001). However, agriculture decision making is a complex problem, as this requires extensive and hard number crunching.

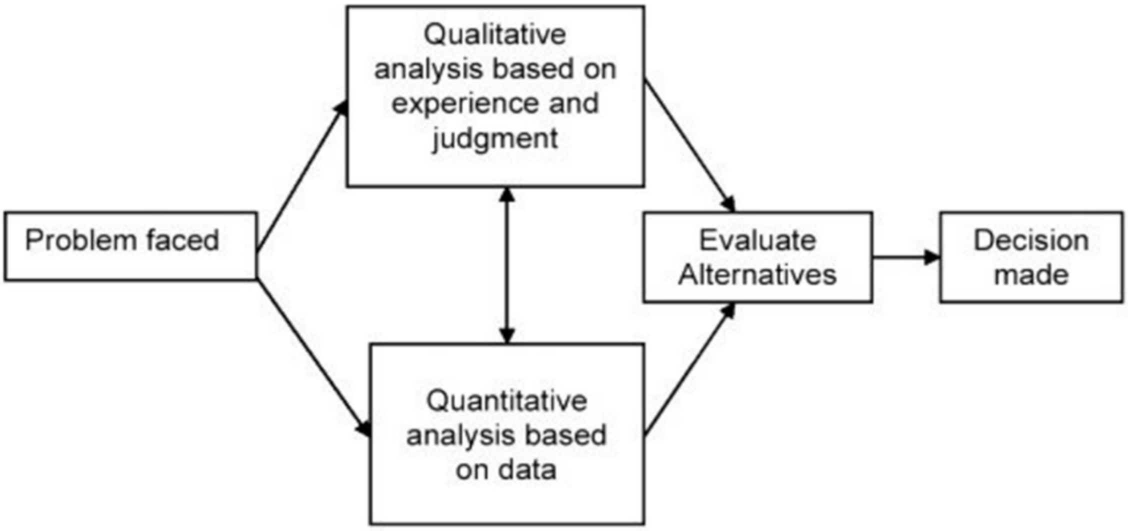

Qualitative analysis based on experience and judgment
Problem faced
Evaluate Alternatives
Decision made
Quantitative analysis based on data

Figure-1: The decision making process (Heinemann, 2010)

To address the complexities of large volume of historical data with low selectivity, a data warehouse is used. A Data Warehouse is the main repository, in fact, a corporate memory, of an organization's historical data. Data Warehouse contains the processed, integrated and clean data used by management for decision support. A data analyst can run complex queries and elaborate analysis using the Data. A Data Warehouse can be viewed as a data-driven approach for running complex queries with low selectivity. Prompt availability of information is vital for strategic analysis as well as for decision making. Online Analytical Processing (OLAP) is used as a supporting application to obtain high performance analysis running ad-hoc "queries" using the Data Warehouse.

2.2 Data Quality

Data is the raw material on which a data warehouse runs. If we consider the analogy of water, tap water is suitable for bathing and washing clothes, but is not suitable for drinking or cooking. Similarly, the bottled water is suitable for drinking, but not for intravenous injection. Similarly, data of right quality and quantity is required for desirable results that are motivated by suitable decision support. Figure-2 shows the six main dimensions of data quality (Hichhorn, 2015). Very briefly, validity deals with compliance with requirements such as, application of definitions, consistency over time and consistency with others. Accuracy means is the data accurate enough for intend purposes and use, such as, is there balance between use, cost and effort and how closely the data is to the point of activity and that the accuracy compromises are clear. Timeliness influences decision i.e. how quickly the data was captured after the event and how quickly it is available and how frequently enough. Completeness is matching quality to meet data needs which includes missing data, invalid data and incomplete data.

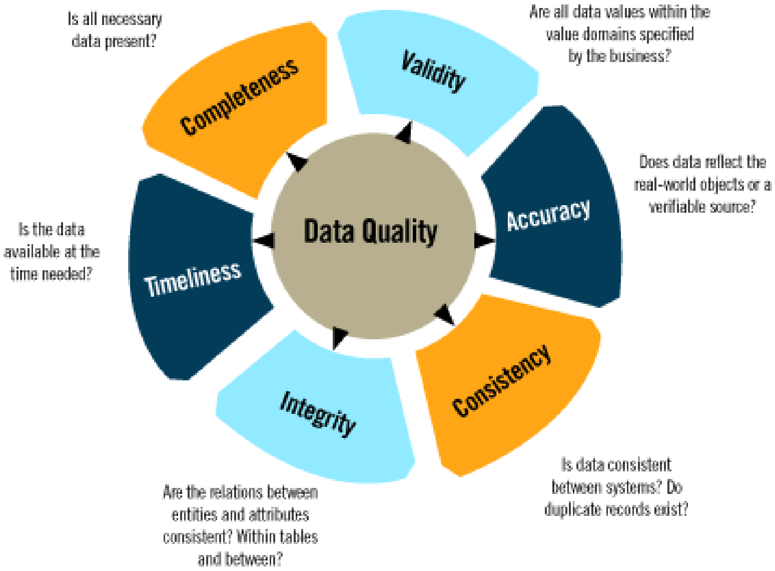

Is all necessary data present?
Completeness
Validity
Are all data values within the value domains specified by the business?
Does data reflect the real-world objects or a verifiable source?
Accuracy
Data Quality
Is the data available at the time needed?
Timeliness
Integrity
Consistency
Are the relations between entities and attributes consistent? Within tables and between?
Is data consistent between systems? Do duplicate records exist?

Figure-2: Data Quality Dimensions (Hichhorn, 2015)

2.3 The Need for ADSS

The agriculture decision-making environment is complex. As per Tarrant (1974), there are three possible methods to the theoretic approaches to agriculture: the first supposes that the physical environment controls agricultural decision-making; the second assumes that uniform producers react in a uniform and rational manner in response to economic circumstances this is called as economic determinism; and the third assumption recognizes that there are additional set of stimuli and effects on agriculture which are neither based on economic nor on physical-environmental factors.

Information Technology (IT) facilitates the smooth running of systems and processes. It is not merely a theoretical discipline but a system for solving problems by using tried and tested techniques. Agro-Informatics i.e. Agriculture + IT is a hybrid of different knowledge-based systems and disciplines that include (but not limited to), agri-sciences, computer sciences, statistics, remote sensing and GIS. IT can enable not only experts and policy makers but also the farmers to utilize the intellectual resources and find the solution of agro-related problems. With the ever-increasing complexity of farming operations both in nature and scope resulting in data generation, the need for prompt and swift flow of information has become important.

3. Related Work

The initial concept of ADSS was presented in (Abdullah et al., 2006). The paper presents a comprehensive discussion of a complete life-cycle implementation of a Pilot Agriculture Extension Data Warehouse. This was followed by query-based data analysis. Some interesting conclusions have been drawn through data mining, using an indigenous clustering technique. Actual cotton pest-scouting data of only 1,500 farmers consisting of about 4,000 data records for 2001-02 of District Multan was processed and used in the pilot project with few forecasted values of weather parameters. The ADSS discussed in the current paper is a full-scale system unlike the pilot project.

NASS (National Agricultural Statistics Service), USDA (U.S. Department of Agriculture) has developed a user-friendly Data Warehouse System that integrates prior survey and census data, and makes the data readily accessible to all NASS employees (Nealon, 2008). Users of the integrated and generalized Data Warehouse are required to navigate only seven tables in a star scheme. The schema consists of the central fact table containing all the survey and census data, and six dimension tables that provide all the necessary metadata to readily access the data. The ADSS developed is different from the NASS Data Warehouse because it uses pest-scouting and Metrological data. In addition to this, ADSS not only provides useful information to the researchers but has also been designed for decision makers and farmers, so that they can get timely information and make suitable decisions based on analysis of historical data.

A NATP (National Agriculture Technology Project) called Integrated National Agricultural Resources Information System (INARIS) was undertaken at IASRI (Indian Agricultural Statistical Research Institute) India. In this project, Central Data Warehouse (CDW) was developed. CDW provided systematic and periodic information to research scientists, planners, decision makers and development agencies in the form of an Online Analytical Processing (OLAP) decision support system (Rai 2007). The ADSS is different from INARIS as the system provides detailed analysis to decision makers and farmers using OLAP as well as query-based applications that provide analysis at District, Tehsil and Markaz Level based on monthly and weekly grain of data.

The National Electronics and Computer Technology Center (NECTEC) in collaboration with the Ministry of Agriculture in Thailand launched an "Agriculture Information Network (AIN)" in response to the information requirements of the agricultural sector (Paiboonrat, 2002). Farmers can get access to the contents through the Internet by themselves or from groups of professionals called "Information Brokers". Although the ultimate beneficiaries of both ADSS and AIN are farmers, but the scope of ADSS is comparatively smaller.

4. Data Extraction and Transformation

ADSS provides decision support by virtue of performance reporting, chart generation, farmer demographic analysis, pest and pesticide analysis, predator and cotton-yield analysis etc. Analysis is performed using historical and latest-recorded data from which projected and derived outcomes are ascertained. Using ADSS processes, the utility of the extracted data is increased by transforming the outcomes into derived attributes and projected results. The major steps of the ADSS workflow are shown in Figure-3.

1. **Data Collection:** The scouts of the Directorate General of Pest Warning (http://www.pestwarning.agripunjab.gov.pk/) visit the target area i.e. South Punjab including District Multan. After interviewing the local farmers, the scouts fill the corresponding information in pest-scouting sheets. The summaries generated from these sheets serve the purpose of informing policy makers where pest hot spots for a particular pest have developed.

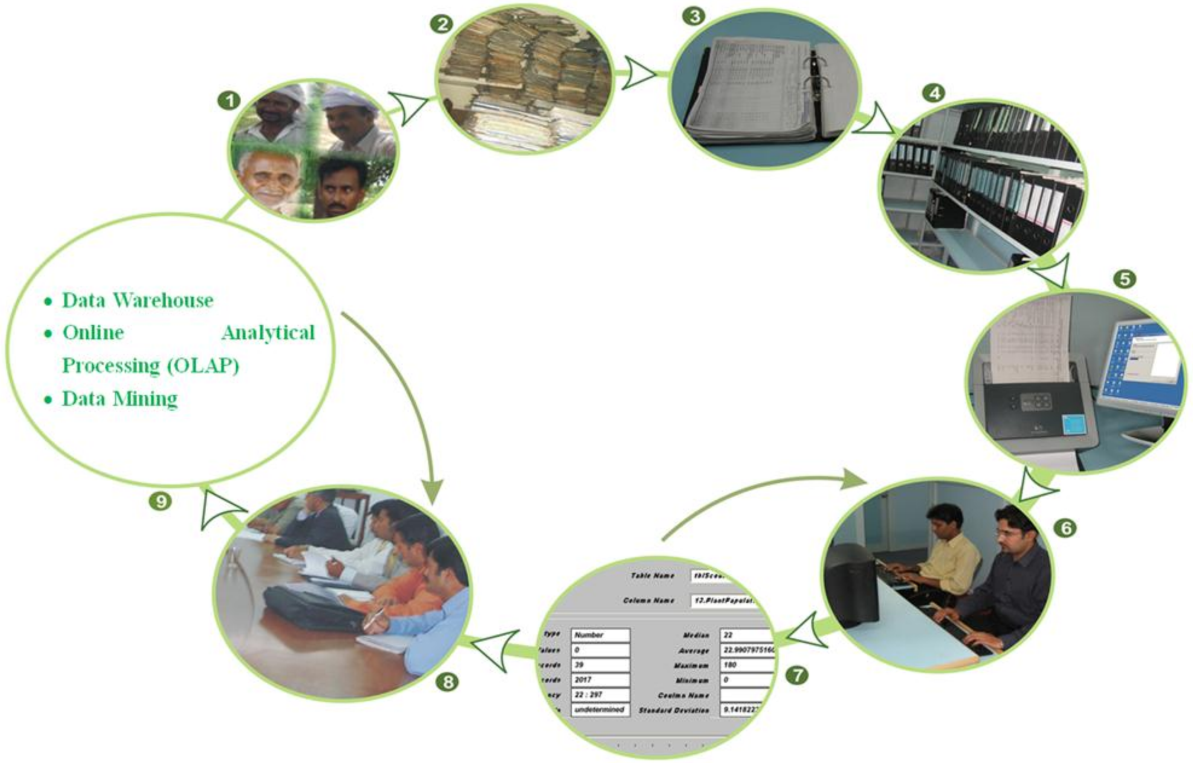


Figure-3. ADSS Work Flow

2. **Archiving of Data Files:** Manually-filled pest scouting sheets are put in files and stored in rooms in different cities of Punjab. This has been the practice for several decades.

3. **Folder Preparation Process:** The sheets at the Directorate of Pest warning are copied and shifted to ADSS premises where these sheets are assigned unique IDs, sorted and separated according to their IDs and subsequently organized into properly labeled folders.

4. **Folder Maintenance:** The folders are maintained in the folder bank for easy retrieval by following a spatial indexing procedure for locating the folder of interest. The folder bank is accessible to data entry personnel who need the pest-scouting sheets for the purpose of scanning, data-entry and error reconciliation.

5. **Scanning of Sheets:** Every pest scouting sheet is scanned and saved as an image in the database. A scanning plan and storage is hierarchy is created prior to the actual scanning. The objective is to ensure that all scanned sheets are just as accessible (actually more) as compared to the original pest-scouting sheets in hard-copy format.

6. **Data Digitization:** A team of Data-entry Operators transform the manually filled data sheets into digitized form i.e. transcribed in a staging database as text/numbers. The process is completed using an indigenous data-entry tool.

7. **Data Cleansing and Transformation:** Digitized data is then analyzed using data profiling tools and necessary transformation performed and then loaded in the data warehouse.

4.1 Data Collection

This section describes in detail the data collection procedure. The process includes: data collection, folder preparation, scanning, data-entry process, data cleansing and data validation.

Pest-Scouting data is collected periodically in the Province of Punjab by the Directorate General of Pest Warning and Quality Control of Pesticides (DGPWQCP) and is recorded on pest-scouting sheets. The scouts from the Directorate General of Pest Warning and Quality Control of Pesticides weekly sample 50 points in each Tehsil of the cotton- growing districts of Punjab. Historically, 3,000 sample points in 60 tehsils of Punjab province are sampled with roughly 150 of these points are from District Multan. It is estimated that until now cotton pest-scouting has resulted in about 3 million records.

The pest-scouting data is recorded on hand-held pads in the field by field scouts, which is subsequently typed or handwritten on pest scouting sheets. Photocopied pest-scouting sheets were acquired from DGPWQCP over a period of several weeks. The data sheet has 27 attributes, such as farmer’s name, date of scout visit, variety of cotton sown, land owned (in acres), date of sowing, plant population, bollworm infestation, incidence of sucking pests, incidence of Cotton Leaf Curl Virus (CLCV), pesticide spray date, etc.

Other than statistical tests of data quality, such as p-value, t-test and ANOVA, a rigorous field-based procedure was adopted for validation of the data recorded on the pest scouting sheets. ADSS team conducted field visits of District Multan. This was meant to learn and observe the data acquisition techniques in the field conditions and to meet the randomly selected farmers in person whose data is used in this study. The ADSS team verified the pest-scouting information after interviewing farmers at different cotton farms. For this purpose, modified pest-scouting sheets similar to the DGPWQCP sheets were used, but with some additional information recorded (farmer demographics etc.). The ADSS team recorded additional data about farmers, such as farmer education, area owned, mobility/transport (car or motor bike), accessibility to TV, radio, computer, internet etc. A total of 36 farmers were visited by the ADSS team in four Tehsils. Furthermore, separate meetings were held with different officers of the Directorate General of Pest Warning and Quality Control of Pesticides (DGPWQCP) in which data collection and data quality processes were discussed.

4.1.1 Process of Folder Preparation

The pest-scouting sheets obtained from DGPWQCP from Multan and Lahore cities were photocopied, labeled with lead pencil and sent back to ADSS premises in the form of packets. This procedure was adopted to ensure that pest-scouting sheets obtained by the ADSS team were in the same order as those of DGPWQCP. The packets were opened and the pest-scouting sheets were punched and organized in folders. The average number of sheets per folder was 350. The folders were prepared methodically so as to make the sheets uniquely identifiable. By way of this process duplicate sheets were eliminated and redundant data was not entered into the database.

Figure-4 illustrates the whole process of the preparation of folder and their storage in the folder/data bank for easy retrieval in the future data-entry purposes. One folder requires approximately 22 man-hours of work to make the folder useable for data-entry into the database.

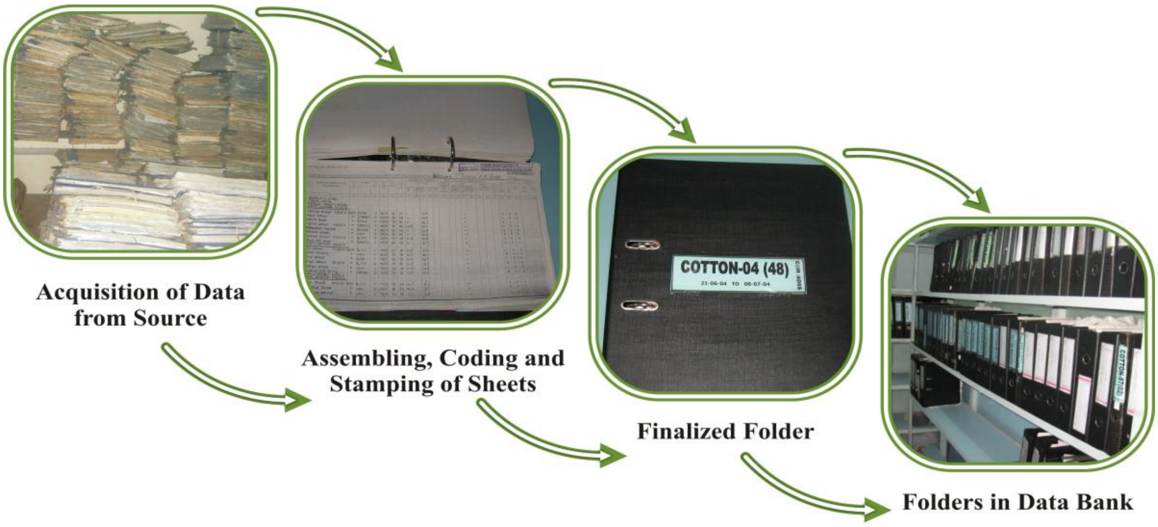
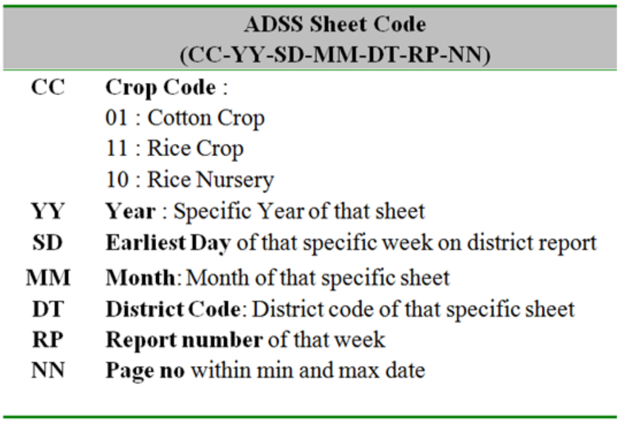

Acquisition of Data from Source
Assembling, Coding and Stamping of Sheets
COTTON-04 (48)
Finalized Folder
Folders in Data Bank

Figure-4: Folder Preparation Process

Description of the components of the ADSS sheet ID used is given in Table-1. After folder coding, the folder is cross-checked for any possible mistake made in the code-allocation process.



ADSS Sheet Code (CC-YY-SD-MM-DT-RP-NN)	
CC	**Crop Code** : 01 : Cotton Crop 11 : Rice Crop 10 : Rice Nursery
YY	**Year** : Specific Year of that sheet
SD	**Earliest Day** of that specific week on district report
MM	**Month**: Month of that specific sheet
DT	**District Code**: District code of that specific sheet
RP	**Report number** of that week
NN	**Page no** within min and max date

Table-1: ADSS Scouting Sheet ID Coding

4.1.2 Data Sheet Archiving

Following folder preparation, the pest-scouting sheets are scanned and saved in the computer. The purpose of this exercise is to keep a digital record of all the data sheets at the ADSS premises to ensure safety against loss due to moisture, termites, fire etc. ADSS used multiple scanners with vertical document feeder allowing fast scanning.

4.1.3 Sheet Scanning

Each ADSS data collection agent scans the sheets from one folder at a time. The sheets from one folder are neither assigned to more than one person, nor are the folders split. This restriction is to ensure that sheets from various folders do not become mixed and that they remain in their original order in the folder undisturbed. Each scanned sheet is then saved as a .jpg file.

4.1.4 Saving the File

The soft folder is named according to the name of the folder in hard form. Tags or separator dates (from the hard copy) are used to create more folders in the parent folder along with one for the summary, which is compiled during the folder preparation. Each of these has a folder name according to the districts scouted during that date (separator). The scanned pest-scouting sheets are saved in the district files. Each image is saved according to the unique ID which has been stamped on the sheet. The division of sheets according to each district ensures that the risks of image mixing in a file are eliminated. Once saved on the hard drive, they are additionally copied on a writable DVD as a backup.

4.1.5 Image Quality

The scanned images are in .jpg format. Pictures in this format take less space as compared to bitmaps of the similar quality. The images are scanned using resolution settings of ‘200dpi’ and paper size is set to ‘Legal’. This translates into size of 1.5 to 1.7 Mb per sheet image.

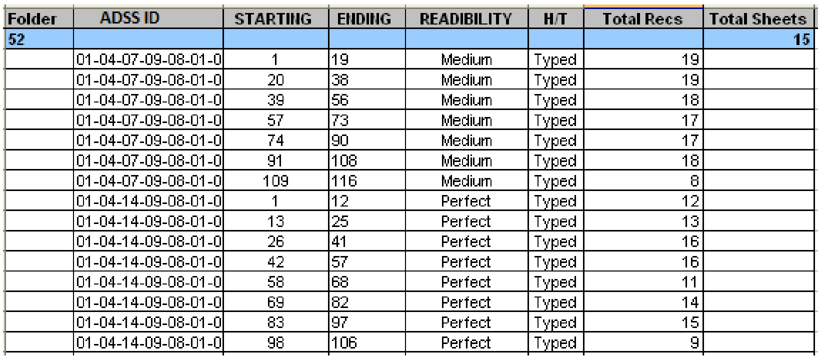
4.2 Process of Data-entry

ADSS uses SET-C (Scouting Entry Tool for Cotton) for digitizing data recorded in the pest-scouting sheets. Benefits of this tool are as follows:

a. SET-C contains a First, Second and Third time data-entry process which assures the quality of data by identifying mismatches found in one record after a second-time entry. The highlighted mismatches are resolved in the reconciliation (third-time entry) stage.
b. Built-in checks prohibit the entry of duplicate records and sheets.
c. Performance, mismatch, reconciliation and summary reports can be generated on daily, weekly and monthly basis.
d. Data is directly stored in the database server.
e. New varieties and pesticides can be added easily.
f. The search option is available in SET-C which makes it possible to search any sheet or record promptly.

After scanning, sheets from different folders are issued to Data-Entry Operators (DEO’s) according to the targeted districts. The following Data-Entry process is followed:

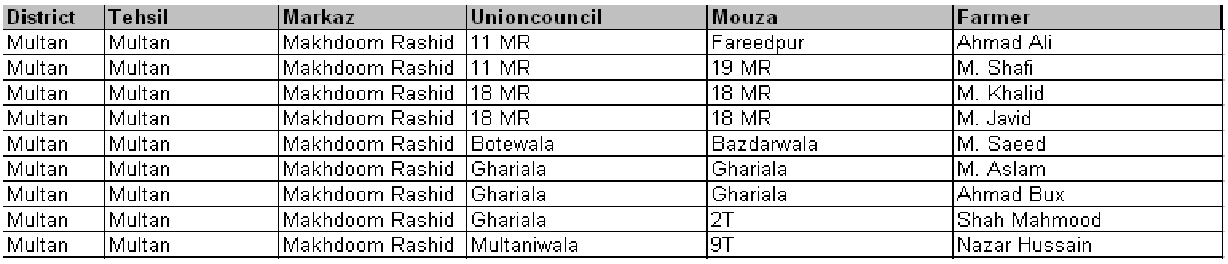
1. Determining which district’s data is to be entered and for which year. Information related to the target district is copied from summaries prepared at folder preparation time (see section 4.1.1) and initially kept in an ‘Excel’ file as shown in Figure-5. This also includes the initial plan of assigning sheets to different data-entry operators.



Folder	ADSS ID	STARTING	ENDING	READIBILITY	H/T	Total Recs	Total Sheets
52							15
	01-04-07-09-08-01-0	1	19	Medium	Typed	19	
	01-04-07-09-08-01-0	20	38	Medium	Typed	19	
	01-04-07-09-08-01-0	39	56	Medium	Typed	18	
	01-04-07-09-08-01-0	57	73	Medium	Typed	17	
	01-04-07-09-08-01-0	74	90	Medium	Typed	17	
	01-04-07-09-08-01-0	91	108	Medium	Typed	18	
	01-04-07-09-08-01-0	109	116	Medium	Typed	8	
	01-04-14-09-08-01-0	1	12	Perfect	Typed	12	
	01-04-14-09-08-01-0	13	25	Perfect	Typed	13	
	01-04-14-09-08-01-0	26	41	Perfect	Typed	16	
	01-04-14-09-08-01-0	42	57	Perfect	Typed	16	
	01-04-14-09-08-01-0	58	68	Perfect	Typed	11	
	01-04-14-09-08-01-0	69	82	Perfect	Typed	14	
	01-04-14-09-08-01-0	83	97	Perfect	Typed	15	
	01-04-14-09-08-01-0	98	106	Perfect	Typed	9	

Figure-5: Multan sheets selected from folder summaries to be entered into the database

2. A list is prepared in ‘Excel’ file (which is later loaded into an SQL database) with the information containing District, Tehsil, Markaz, Union Council, Mouza and Farmer’s Name as shown in Figure-6:



District	Tehsil	Markaz	Unioncouncil	Mouza	Farmer
Multan	Multan	Makhdoom Rashid	11 MR	Fareedpur	Ahmad Ali
Multan	Multan	Makhdoom Rashid	11 MR	19 MR	M. Shafi
Multan	Multan	Makhdoom Rashid	18 MR	18 MR	M. Khalid
Multan	Multan	Makhdoom Rashid	18 MR	18 MR	M. Javid
Multan	Multan	Makhdoom Rashid	Botewala	Bazdarwala	M. Saeed
Multan	Multan	Makhdoom Rashid	Ghariala	Ghariala	M. Aslam
Multan	Multan	Makhdoom Rashid	Ghariala	Ghariala	Ahmad Bux
Multan	Multan	Makhdoom Rashid	Ghariala	2T	Shah Mahmood
Multan	Multan	Makhdoom Rashid	Multaniwala	9T	Nazar Hussain

Figure-6: Demographics list of Multan sheets that are uploaded into the database

3. The purpose of this list is to maintain consistency regarding farmers’ names and their demographics. Once the list is prepared, it is entered into the SET-Cotton system tables by the DBA. The Data-Entry Operators can now select the appropriate entry from a drop-down menu. One person prepares the list by looking at the pest-scouting sheets. This method is time consuming to implement, but the cost is effectively amortized across all entries made.

4. The DEO Team Lead adds pest-scouting sheets one by one in SET-C for Data-entry. The ‘Excel’ file prepared in Step 1 greatly reduces the time required for this process as all required sheets and their relevant information is available at one location.

5. After adding sheets in SET-C, the DEO Lead assigns these sheets to Data-entry Operators for first-time data-entry. The assignment of sheets is done according to the list prepared during Step 1.

6. Data-Entry Operators use the form containing pest-scouting sheet details for data-entry. The demographics data is entered from a drop down menu using the list created in Step 3.

7. After the first-time data-entry, the DEO Lead receives that particular sheet and assigns the sheet to another DEO for second-time data-entry. The same process is repeated when the second-time data-entry is performed and is ready for reconciliation. Only the records of a sheet are visible to the third DEO (to whom the sheet was assigned) which do not match for

first and second time data-entry, these mismatches are checked and corrected at the time of reconciliation. The reconciliation process is the final step in data-entry. The DEO lead finalizes the reconciled data sheets in SET-C. After reconciliation the data is ready to be cleansed.

4.3 Data Quality Management

The process of data quality management adopted in ADSS is as per Figure-7. Here SME is the subject matter expert. Note that the self-explanatory process is iterative and the more the number of iterations, the higher will be the data quality. But one must keep in mind the cost-quality trade-off i.e. unless one is careful, in the extreme case the cost of achieving high data quality will increase exponentially with little increase in actual data quality.

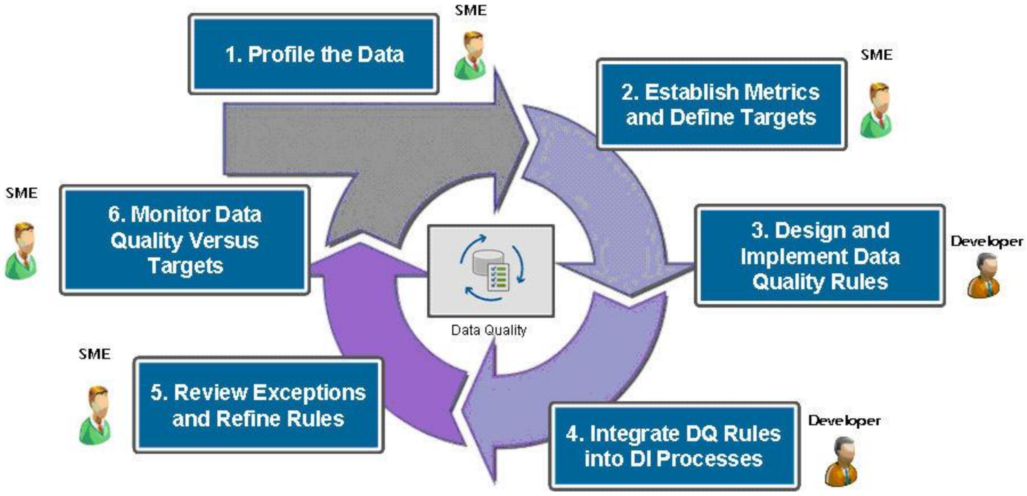

SME
1. Profile the Data
SME
2. Establish Metrics and Define Targets
3. Design and Implement Data Quality Rules
Developer
4. Integrate DQ Rules into DI Processes
Developer
5. Review Exceptions and Refine Rules
SME
6. Monitor Data Quality Versus Targets
SME
Data Quality

Figure-: Data Quality Methodology (DeLua, 2009)

The pest-scouting sheets contain non-standardized as well as erroneous data recorded by pest scouts during surveys. Moreover, errors could have been made during the data-entry process due to the poor quality of photocopied pest-scouting sheets. The end result is a dataset which is dirty and not ready for scientific analysis. Data standardization and cleansing procedures are performed to address these problems.

Problems encountered during the discovery of errors and their resolutions are of many types. To avoid confusion and maintain timely delivery of quality data, the entire process is divided into well-defined parts which are performed sequentially (For data transformation details see section 4.4). The procedures used to cleanse and standardize the data are described in the following sections.

The indigenous prototype Data Profiling Tool identifies the business rule violations, trends in data, null values and duplication. These features are utilized to assess data quality and help cleanse the pest-scouting data. The most commonly used features of this tool are ‘detailed profiling’ and ‘summarized profiling’ which show unique value distribution, null values and prevalent trends.

Other features of this tool include 'House Holding' (to identify duplication), Cleansing (loading standard values against non-standard data) and Data Quality Assessment (shows a graph to describe quality of data).

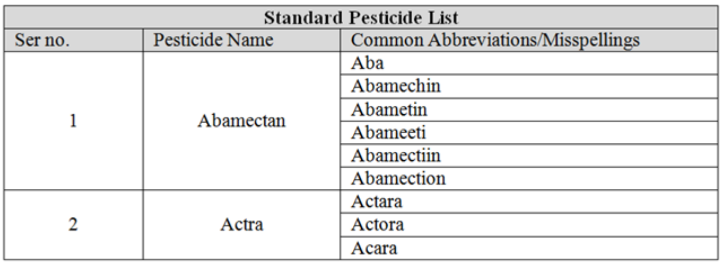
The selected columns are checked by the profiling tool for unique value distributions (for standardization as well as quality assessment purposes) and trends. A unique number of instances, listed in alphabetical order are especially useful in determining misspellings in virus incidence, pesticides, units and location columns. The trends and errors are highlighted and recorded in the quality assessment report. This report contains quality assessment graphs for selected columns, along with general comments and trends in data.

4.3.1 Data Standardization

Erroneous names and misspellings are particular problems for the Pesticide Names', Tehsil and Markaz names, Variety names and Units' columns. The profiling tool and SQL queries are used to find errors in these columns. These problems and their solutions are documented after consultation with agriculture expert.

Example-1

This example shows data standardization for pesticide names. A sample of standard pesticide names is given in Figure-7. It can be seen that the pesticide 'Abamechin' contains six abbreviations and is misspelled on the pest-scouting sheets. These values were converted to one standard name with the help of the data profiling tool by creating a column with the correct value.



Standard Pesticide List		
Ser no.	Pesticide Name	Common Abbreviations/Misspellings
1	Abamectan	Aba
		Abamechin
		Abametin
		Abameeti
		Abamectiin
		Abamection
2	Actra	Actara
		Actora
		Acara

Figure-7: Standard names for pesticides

The same procedure is used for other columns where multiple values of the same name may exist. The same procedure is followed for Variety column. For example the variety NIAB 111 is also commonly written as N-111in the pest scouting sheets.

4.3.2 Data Cleansing

Validation checks on SET-C restrain certain types of data to be entered during the data-entry process. This is necessary to decrease errors at the time of data entry. These errors are recorded in the Data quality document (excluding location columns). Once standardization has been performed, further errors (if any) are removed using the data profiling tool. All these errors (and their resolution in later stages) are recorded in the 'Quality Assessment Report' for future reference and resolution of errors.

4.3.2.1 Dose Unit Errors

No unit has been assigned to the dose of pesticide used i.e. ml or gm. This problem arises when the unit is not given with the pesticide name on the pest-scouting sheet. The problem is solved by using an update query which assigns a unit value where the dose column is not null. Entries in the Data quality document also point out unusual figures, for example numerical values like 1 or 2 are occasionally found in pesticide dose columns due to errors made during data-entry. The entries are shown to the agriculture specialist who provides the solution. Here, entries must be inscribed and corrected manually in the database.

Example-2

This example shows a query that is used for mass updation in instances where units are missing.

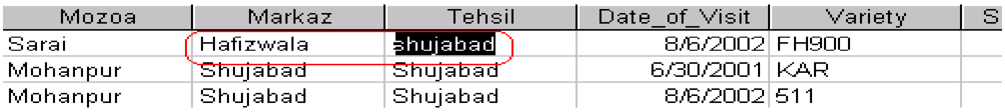
> *Update combined set spray1unit = 'gm'*
> *where spray1pesticide IN ('Actra',* 'Buprofizen', *'Crown', 'Getred', 'Imidacloprid', 'Lanate', 'Larvan', 'Pride', 'Thiodicarb', 'Imicon', 'Pestidor')*
> *AND spray1dosage is not null*

4.3.2.2 Wrong Location Name

This situation occurs when one markaz has erroneously been assigned to more than one district (Markaz is a sub-region in a District). These mistakes are due to ambiguous entries on the pest-scouting sheets. This data quality problem is handled during standardization.

Example-3

This example explains how incorrect location names are corrected in the database. Figure-8 shows a problem in the pest-scouting data of the year 2001 for Tehsil Shujabad. Hafizwala is shown as one of the Markaz in this Tehsil, where in reality only one Markaz (Shujabad) is located. If the errors are minor, they are corrected manually; otherwise, SQL queries are used.



Mozoa	Markaz	Tehsil	Date_of_Visit	Variety	S
Sarai	Hafizwala	shujabad	8/6/2002	FH900	
Mohanpur	Shujabad	Shujabad	6/30/2001	KAR	
Mohanpur	Shujabad	Shujabad	8/6/2002	511	

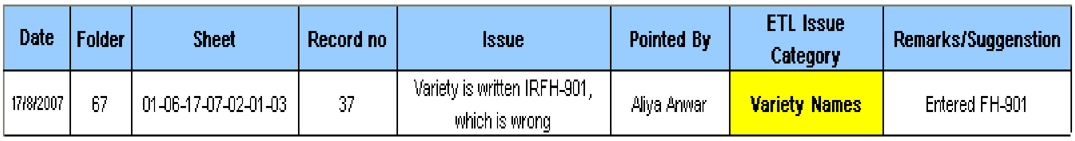
Figure-8: Incorrect geographic entry in the database

4.3.2.3 Variety Names

Improper variety names are sometimes seen on the pest-scouting sheets. This is usually due to hasty entries made by scouts or illegible handwriting.

Example-4

This example shows how incorrect crop variety names are corrected during the cleansing phase. These entries are almost always corrected at the time of data-entry with the consultation of the agriculture specialist. Unresolved errors at the time of data-entry are resolved in the cleansing phase. SET-C uses a drop-down list for the variety column and any variety name not found in that list is dealt with as per the method described. Figure-9 shows an example of this problem and its solution in the Data quality document. An incorrect entry of IRFH-901 was entered correctly as FH-901.



Date	Folder	Sheet	Record no	Issue	Pointed By	ETL Issue Category	Remarks/Suggenstion
17/8/2007	67	01-06-17-07-02-01-03	37	Variety is written IRFH-901, which is wrong	Aliya Anwar	Variety Names	Entered FH-901

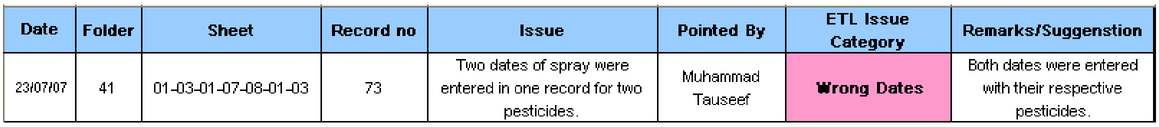
Figure-9: Solution for incorrect variety name entry

4.3.2.4 Issues of wrong dates

This is one of the most common problems encountered in pest-scouting data. Cross checking is required with the original pest-scouting sheets in order to fix these problems. Some of these problems have to be solved at entry time and recorded in the Data quality document.

Example-5

This example explains issues related to date columns and their resolution. Figure-10 shows an entry that was confusing as two dates were used for one pesticide. As the later date was not possible (due to the given visit date), only the first date was used. This issue was solved at the time of data-entry.



Date	Folder	Sheet	Record no	Issue	Pointed By	ETL Issue Category	Remarks/Suggenstion
23/07/07	41	01-03-01-07-08-01-03	73	Two dates of spray were entered in one record for two pesticides.	Muhammad Tauseef	Wrong Dates	Both dates were entered with their respective pesticides.

Figure-10: Solution for wrong date issues

4.3.3 Data Cleansing Using Business Rules

Business rules (total 22 business rules) have been developed as part of ADSS for all attributes of the pest-scouting sheets to find errors or unusual entries. Some of these rules are described as follows:

Rule-1: Plant Population: Cannot be zero. If so then these might have been left out by scouts. Values can be from 20,000 to 75,000.
Rule-2: Whitefly Adult: 1 to 10 (ETL = 5)
Rule -3: Pesticide Spray Date: Must be later then sowing date and earlier the visit date.
Rule-4: Area: Cannot be zero
Rule-5: Variety: Cannot be zero. Must have a prefix (variety type) before number
Rule-6: CLCV Incidence: Cannot be greater then 100.
Rule-7: Predators: Can be up to 50.

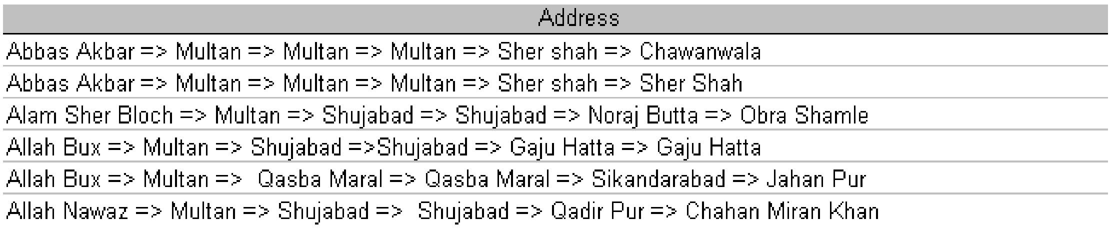
Note that here ETL (Economic Threshold Level) is the pest population beyond which it is economical to use pesticides. The errors found by implementing these rules must be cleansed by manually cross-checking with the original pest-scouting sheets. The record of any changes made is written in the Quality Assessment Report.

4.4 Data Transformation

The data is received by the Data Quality Assurance Manager in an ‘MS Access’ database in a denormalized, flat-file form. The main reasons for this is that the prototype profiler works only on MS Access and cleansing process becomes much easier when dealing with data in flat file form. During the data transformation phase, different non-standard column values are converted into a standardized format. Some important conversions which take place are as follows:

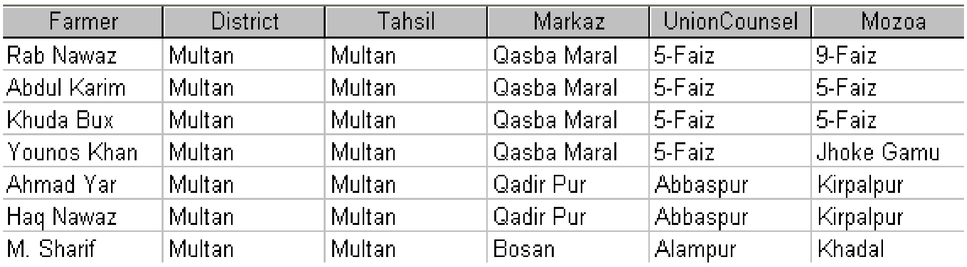
4.4.1 Transforming Farmer Demographics Column

Data-entry for farmer demographics is done through a list prepared before actual data-entry begins. The farmer demographics details are entered in the SET-C database to ensure that spelling mistakes are kept to a minimum and to decrease data-entry time. Data in this form has to be divided into separate columns. Transformation of the farmer demographics list, as entered by DEO’s takes place, and one column is converted into 6 distinct columns (Figure-11 and Figure-12) i.e. 1:M transformation.



Address
Abbas Akbar => Multan => Multan => Multan => Sher shah => Chawanwala
Abbas Akbar => Multan => Multan => Multan => Sher shah => Sher Shah
Alam Sher Bloch => Multan => Shujabad => Shujabad => Noraj Butta => Obra Shamle
Allah Bux => Multan => Shujabad =>Shujabad => Gaju Hatta => Gaju Hatta
Allah Bux => Multan => Qasba Maral => Qasba Maral => Sikandarabad => Jahan Pur
Allah Nawaz => Multan => Shujabad => Shujabad => Qadir Pur => Chahan Miran Khan

Figure-11: Actual form of data as entered by DEOs in the SQL server



Farmer	District	Tahsil	Markaz	UnionCounsel	Mozoa
Rab Nawaz	Multan	Multan	Qasba Maral	5-Faiz	9-Faiz
Abdul Karim	Multan	Multan	Qasba Maral	5-Faiz	5-Faiz
Khuda Bux	Multan	Multan	Qasba Maral	5-Faiz	5-Faiz
Younos Khan	Multan	Multan	Qasba Maral	5-Faiz	Jhoke Gamu
Ahmad Yar	Multan	Multan	Qadir Pur	Abbaspur	Kirpalpur
Haq Nawaz	Multan	Multan	Qadir Pur	Abbaspur	Kirpalpur
M. Sharif	Multan	Multan	Bosan	Alampur	Khadal

Figure-12: Transformed form of the data as corrected by the DBA

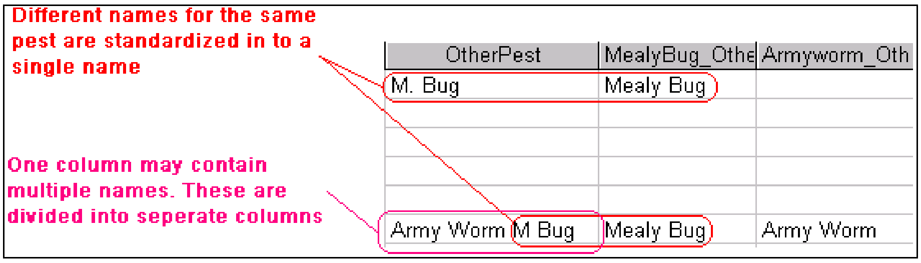
4.4.2 Plant Height Problems

Cm and Inch values are used for recording plant height on the pest-scouting sheets. All inch values are converted to cm by multiplying the value by 2.5 and thereby replacing inches with cm. This is done by an update SQL query. (The use of a distinct query or data profiling tool is necessary to find which kinds of units are being used for one column) i.e. 1:1 transformation.

```
update Combined set plantheight = plantheight*2.5 , plantheightunit = 'cm'
where plantheightunit = 'inch';
```

4.4.3 Separation of other pests

The pest-scouting sheets contain only one column for ‘Other Pests’ and might contain more than one pest. Original entries are divided into separate columns, with standard pest names. Figure-13 shows an example of such a step in the database.



Different names for the same pest are standardized in to a single name

One column may contain multiple names. These are divided into seperate columns

OtherPest	MealyBug_Othe	Armyworm_Oth
M. Bug	Mealy Bug	
Army Worm M Bug	Mealy Bug	Army Worm

Figure-13: Transforming ambiguous entries to clearly readable names

5. Conclusions

Although DSS have been around for a while, and financial or telecommunication data warehouses are not something new, however, an agriculture data warehouse is relatively a new entrant in the domain of DSS. In the absence of an MIS system for agriculture data, managing the corresponding data is a challenge. Therefore, processes and procedures are presented in this paper demonstrating how to manage the pest scouting data. Data quality or the lack of it is an important and time consuming issue in traditional data warehouses, and for number of reasons, this being of higher complexity for an agriculture data warehouse. For agriculture pest scouting data, quality is a hard problem for a number of reasons, such as absence of an online transaction processing system, variability and diversity in the pesticide names, pests, plant viruses and crop varieties. In this paper, we have discussed the data management and data quality problems associated with agriculture pest scouting data, and also given the processes and techniques to fix or reduce the impact of these problems.

References

Abdullah, A., & Hussain, A. (2006). Data mining a new pilot agriculture extension data warehouse. *Journal of Research and Practice in Information Technology, 38*(3), 229-250.

Anonymoushttp://www.statista.com/statistics/263055/cotton-production-worldwide-by-top-countries/ (14 Nov. 2015)

Davidson, A. P. (2000). Soil salinity, a major constraint to irrigated agriculture in the Punjab region of Pakistan: Contributing factors and strategies for amelioration. *American journal of alternative agriculture, 15*(04), 154-159.

Davidson, A. P., Ahmad, M., & Ali, T. (2001). Dilemmas of agricultural extension in Pakistan: Food for thought. Overseas development institute (ODI). Agricultural research & extension network (AgREN).

Gadi Hichhorn, 3 Reasons Why Data Quality Should Be Your Top Priority This Year (2014), http://www.realisedatasystems.com/3-reasons-why-data-quality-should-be-your-top-priority-this-year/ (14 Nov. 2015)

Heinemann, P. H. (2010). Decision support system for food and agriculture. System analysis and modelling in food and agriculture, Encyclopedia of life support system (EOLSS).

Iftikhar, N., Liu, X., & Nordbjerg, F. E. (2015). Relational-Based Sensor Data Cleansing.

In New Trends in Databases and Information Systems (pp. 108-118). Springer International Publishing.

Julianna DeLua, Data quality can put your organization on the defense or offense depending on how well you manage it (2009), https://tdwi.org/Articles/2009/01/28/Data-Quality-Management-Getting-More-from-Your-BI-Investment.aspx?Page=1 (14 Nov. 2015)

Nealon, J., &Yost, M. (2008): Easy and fast data access for everyone. National Agricultural Statistics Services, U.S. Department of Agriculture.

Paiboonrat, P. (2002). IT for Rural Agriculture and Rural Development. Socialist GIS and Spatial Database Section. National Electronics and Computer Technology Center (NECTEC), Bangkok, Thailand.

Power, D. J. (2001, June). Supporting decision-makers: An expanded framework. In e-Proceedings Informing Science Conference, Krakow, Poland(pp. 431-436).

Rai, A. (2009). Data warehouse and its Applications in Agriculture. Indian Agricultural Statistics Research Institute Library Avenue, New Delhi, 175-183.

Tarrant, J. R. (1974). Agricultural geography, 78-82.

Copyright Disclaimer