Using Audit Trail data to move from a Black Box to a Transparent Data Collection Process

Jacqueline Hunt, Central Statistics Office, Ireland

1. Abstract

CSO is currently transforming the management and administration of all its household surveys. Part of this transformation is the introduction of CATI for the LFS waves 2-5. The telephone interviewing is being outsourced and all CATI interviews are being conducted by a third party service provider. The contract negotiated by the office for the service is based on the duration of calls. Early testing with the call centre revealed using timestamps embedded in the questionnaire as a method of generating interview durations did not provide the accuracy required for contract purposes. This led us to look at an alternative method to calculate the interview durations. Our previous use of Audit Trails was limited to using the data to recreate corrupted interviews and for ad-hoc analysis to investigate incidents of unusual interviewing behaviour. The third party contract requirements led us to investigate how we could use the audit trail data to extract more accurate interview durations. As part of the design process we examined other potential uses of the Audit Trail data, such as creating key process indicators that can be used to monitor and improve the data collection process. This paper will look at our design approach and the solution that has been implemented as part of the Household Transformation project.

2. Introduction

Paradata is defined as data about the data collection process that can be used to manage and improve data collection processes. A particular type of paradata, Audit Trail data, is generated as part of the computerised interviewing process. Audit trail files are key logging files that record every keystroke made by an interviewer during the course of an interview. The files created are large, semi-structured text files that are difficult and time consuming to process, however, the data they contain can be very valuable. Finding a way to convert the audit trail data into a useable format that can be searched and queried to provide insights into the data collection process is desirable for many organisations. As part of the CSO Household Survey's Transformation Project a new datamodel structure was developed and implemented for the audit trail data that makes it easier for users to extract information that can be used to measure and monitor the data collection processes and interviewer performance. Access to this data, moves the office from a 'black box' data collection process to one with complete transparency that can be measured and improved.

This paper describes a solution that allows the integration of the audit trail data with other household paradata to provide full transparency across CSO's household surveys field and call centre operations. Finding a solution that converts all the data available within the audit trail files and storing it where it can to be used with other paradata maximises the business value that can be achieved from the data. This project resulted in the creation of new database tables for the audit trail data, with an accompanying ETL (Extract, Transform and Load) process, to parse the data into the new format. Thereby, providing a new structure for the audit trail data; making it easier to work with the data, and easier to extract information that can be used to inform many aspects of the data collection processes.

3. Paradata Benefits

Audit trail files are automatically generated key stroke logging files that are created as part of a computer assisted data collection process, such as face to face interviewing using an electronic questionnaire, telephone interviewing or self-interviewing via the web. The Audit trail data can provide information about how interviewers complete and navigate through the electronic survey

instrument. The data is collected at the time of the interview with the survey data and without interfering with the interview process (Snijkers & Morren, 2010). The audit trail data can be used to provide transparency of the interviewing processes; making it possible to improve the management and quality of the data collection processes. Unfortunately, the audit trail files that are generated are large semi-structured text files that are difficult to process. It is believed that survey paradata can be used to provide a wealth of information about the interview process but it is only as useful as the ease with which the data can be accessed, manipulated, and analysed (Devonshire, et al., 2012). The audit trail files in their natural state are difficult to work with and need to be converted into a different format to make them more usable. There is a consensus in available literature on the subject that the audit trail files contain a wealth of information that can provide valuable insights into the data collection processes for instrument diagnostics, survey methodologists and interviewer managers.

Table 1. Documented Uses of Audit Trail Data

Use of Audit Trail Data
• Better understanding of the data collection processes
• Monitor data collection and identify problems in a timely fashion
• Produce quality control measures
• Inform survey design decisions; use to evaluate survey instruments, alleviate response burden and look for potential difficult areas in the questionnaire
• Minimise survey error and improve the data collection process
• Monitor non response bias and correlation of likelihood of participation
• Review interviewer effect on measurement error using question timings, keystrokes, common suspension points, interviewer response to an error check.
• Look for patterns of interviewer actions associated with fatigue, use the results to improve recruitment, training and supervision of interviewers.
• Manage interviewer performance – monitor date/time stamps are in chronological order
• Examine keystrokes to determine fields with comments, fields associated with backup actions, fields with error checks

4. Problem Evolution

The Household Surveys Development (HSD) project objective is consolidation and automation of routine tasks across all household surveys to deliver efficiencies and improve the processes associated with the data collection operations. A major component of this project is the introduction of a third party contractor to conduct telephone interviews for the repeat interviews of one of the office's flagship household surveys, the Labour Force Survey (LFS). Traditionally, CSO household surveys have all been conducted as Computer Assisted Personal Interviews (CAPI), also known as 'face to face' interviews. Telephone interviewing is being introduced for repeat waves of the LFS to free up capacity among the permanent field staff to make it easier to conduct new surveys. Until now to introduce a new survey, a new survey administration section was required to manage the survey and a new field force had to be recruited, trained and managed for the duration of the survey. Introducing telephone interviews for repeat waves of the LFS will increase the capacity among the permanent interviewers to undertake new surveys as they arise.

The contract payments negotiated with the call centre service provider is based on the duration of complete interviews. Call centre pricing models can vary; examples include costs per call, per agent hour and all inclusive rates per campaign. Usually the trade-off based on the pricing model is

between the quantity and quality of calls (Bucher, 2013). The model chosen should reflect the call outcome objectives that are desired. For CSO, the objective is good quality interviews and respondents' commitment to continue to participate in the longitudinal aspect of the survey. Building a rapport with the respondents to encourage future participation in the survey and the likelihood of a positive response the next time they are called is more important than conducting a large quantity of calls within a specific timeframe. Negotiating a pricing model based on interview durations was chosen to focus the call centre and their agents on the interview quality rather than the volume of calls per hour.

The original business problem emerged during the initial test phases with the call centre when it proved difficult to extract accurate interview durations from the questionnaire. Until now the duration of interviews was not an essential requirement for the field interview operations as it was not a factor in the field interviewer remuneration. The interview duration was previously calculated using time stamps that have been inserted into the electronic survey instruments at predetermined questions. It was accepted that the time stamps provided an indication of the interview duration only as the timestamps built into the questionnaire can prove unreliable for a number of reasons. For example, the multiple routes and exit options which can result in no end-time being recorded; in addition, an interviewer may open and close the case a number of times, this may be to preview the case before a call or correct text values after a call, a high number of sessions per case can result in mis-matched start and end times. The contract with the call centre highlighted the deficiencies using the timestamps to calculate accurate interview durations. The time stamps did not provide the accuracy required to calculate contract payments. The test case results included incidents of cases that did not match the third party recorded times and incidents where it was not possible to calculate interview durations due to missing end-times.

An alternative paradata source is available that records all interviewer activity with the electronic questionnaire during the interview; including the start and end times of every interview session. The audit trail file is a key logging file that is generated automatically during the interview process. The audit trail data captures every key stroke activated during an interview session. As a result a large volume of data is generated for each interview session however, the volume of data generated leads to a more accurate representation of the interviews. This data is in a semi-structured format which means it cannot be queried or searched in its original form. Current uses of the audit trail data had been limited to recreating interviews if they've been corrupted or investigating interviewer performance if there is a suspicion of inappropriate behaviour. Both of these tasks were manual, laborious, complex ad-hoc processes. However, the options available to resolve the call centre contract problem were either to develop a more robust way of recording time stamps in the questionnaire or finding a solution to use the audit trail data. The difficulties associated with the timestamps focused attention on the possibility of finding a more efficient way to use the audit trail data to extract the duration information required.

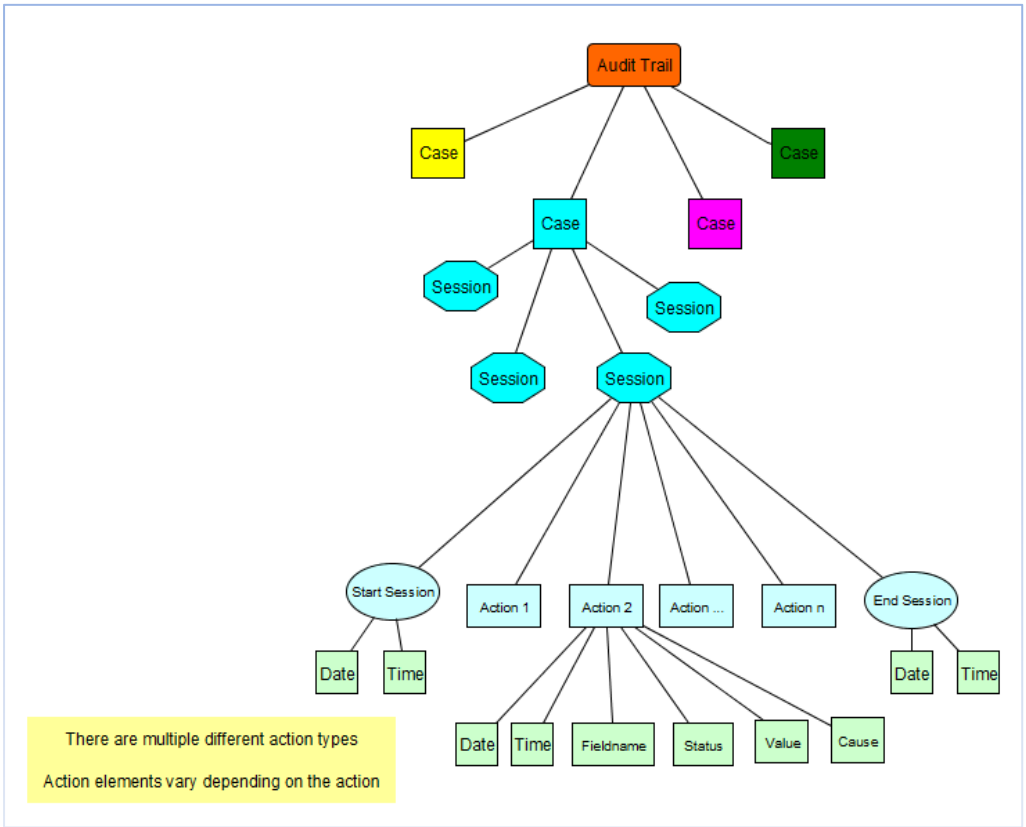
5. Solution

The primary objective was to build a set of database tables and a parsing tool that could be used to restructure the audit trail data to enable easy extraction of the interview durations. Semi-structured data is characterised by a lack of a fixed schema and record layout, although typically the data has some implicit structure. The audit trail data is further complicated because the number of actions taken by the interviewers are not fixed in advance resulting in a variable number of observations per interview session (Olson & Parkhurst, 2013). In addition, each record can have a different structure and length. The goal of working with this data, presenting and querying it is impaired by its original structure. The critical problem to resolve is the discovery of the structure implicit in the raw data and, subsequently, recasting the data into a structure that is easier to work with (Nestorov, et al., 1998).

Patterns in the data were used to map and identify data elements, see Figure 1. Depending on how the audit trail is generated it can contain many case files. Each case will contain one or more interview, browsing and/or editing sessions. Each session will have a start and end action with an associated date and time. Each session usually has a number of actions; the number of actions is a variable

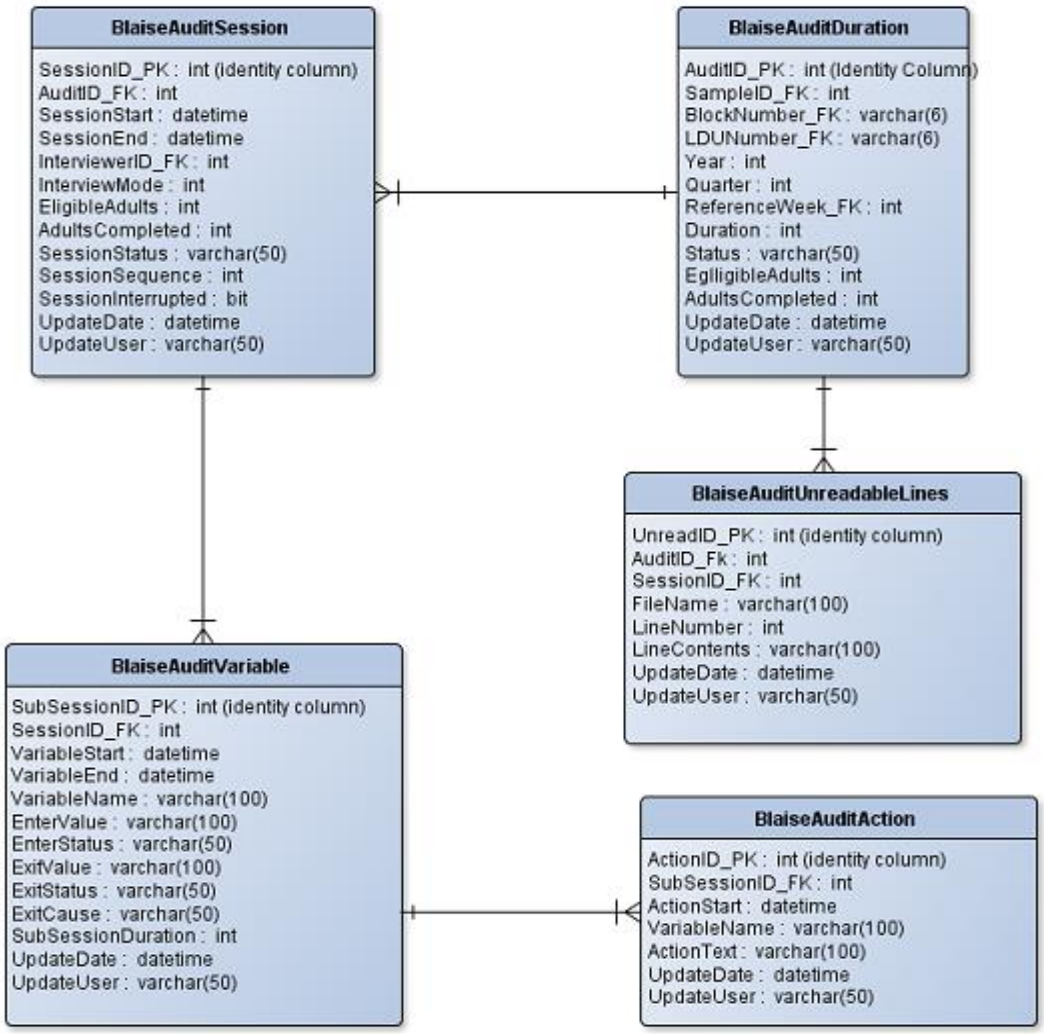
number completely dependent on what happens during an interview or editing session. There are a number of configuration options available that control how the files are generated. The current on-going household surveys generate an audit trail file by interviewer spanning a survey period. In this scenario there are multiple cases in each file, with corresponding lower level details. Every time a case is accessed by an interviewer the details are appended to the audit trail file. To facilitate transferring the data to the office the files are purged after a successful data sync with the office, to help manage the size of the files over the survey period. For our new developments the survey case files are being structured differently, each audit trail file will hold the details of one case only but there can still be multiple sessions and corresponding lower level details. The added complexity will also be that there may be different interviewers as the audit trail file is now part of the total case object package and will be kept with the case details while the case is active in the survey.

Figure 1. Audit Trail Data Graph



A set of hierarchical database tables were designed around the data elements to include a session table, containing details of each session; a duration table, to keep track of the overall time that the case was accessed; a variable table, to record each variable that was acceessed and an action table to hold all actions associated with the variables. Some additional aggregrate fields have been added to allow for easy access to the duration of an interview. A misc. table was also created to hold details of lines that could not be processed. An analysis of the records been added to this table identified them as being caused by an interrupted session where no end time was being recorded. Consulting with the users it was decided to manage these records by identifying them as interrupted records and adding the last time recorded to the record ensuring a total duration could still be calculated for all sessions and that interrupted sessions could be identified.

Figure 2. Audit Trail Database Implementation



An ETL process written in Java was used to populate the data tables. Each audit trail record is searched to find an identifying record attribute and then written to the corresponding table depending on the attribute found e.g. 'Open Session', 'Enter Field' etc. Each record added to the table is given a unique identifier that maintains the sequence of records from the original audit trail data. The final datamodel is not completely normalised but this was not the goal. The data is not required to support a transactional process rather the goal is to make it easy for users to query and search the data. Fully normalised data can be easier to import and update but it can be harder to get the data out as multiple tables and join paths can make queries less efficient and harder to code correctly (Corr & Stagnitto, 2012).

This project has resulted in a new data model created for the audit trail data that can be used to extract accurate interview durations for CATI and CAPI interviews. A new ETL process to handle the data import has been developed and tested. A new CATI Interview Duration report has been developed. The audit trail data parsing process has been added to the overnight case processing job ensuring the latest audit trail data is available to the business as it is received. The data quality and accuracy of the audit trail data has been verified and we can be confident that we are producing accurate interview durations for the contract. This is a good resolution of the original problem that meets all the requirements of the call centre contract manager; who can be confident in their ability to handle the call centre contracts based on accurate interview durations.

The audit trail data can now be used with the other para data such as the Call History data to provide complete transparency across the household survey data collection operations. This moves the office from a scenario where there were a lot of unknowns about the data collection in the field to having complete transparency across the data collection process for both modes; 'face to face' and telephone. The data provides a wealth of information to fully understand how the data collection process is operating. Data from the audit trails can be used to track days and times of interviews, monitor interview durations and review interviewer performance. The solution can be used to deliver automated reports to the business users on the field and telephone interviewing operations. The overall result should be an improvement in the management capabilities for household survey data collection through having access to detailed information on what actually happens in the field.

6. Potential Use of the Data

The audit trail data can be used to improve a number of the household survey operational and design processes ensuring they conform to the highest standards and best practices. The project identified opportunities to use the audit trail data to support process improvements for data collection processes via managing and monitoring key performance indicators. The audit trail data can be used to provide performance indicators to monitor and improve questionnaire design and testing processes, management of field and call centre operations and interviewer performance thus, leading to improvements in the overall data quality. Working with business stakeholders we were able to identify the following types of reports that could be created from the audit trail data to assist with managing the data collection operations.

Table 2. Reports Based on Audit Trail Data

Report Name	Description
Interview Duration	Interview durations viewed by session, mode, interviewer
No. of Sessions	Number of session to complete an interview - Analysis by mode, interview, duration of session
Interview Activity	Interview activity by day of the week and time of the day
Module Duration	Module duration – specify start and end question to calculate duration for a section of the interview
Interviewer Performance	Compare interview durations by interviewer to identify outliers (possible indication of performance issue)
Aggregate Performance	View durations, active days and time across multiple surveys by interviewer - report that integrates audit trails from all four surveys so that interviewer workload and working patterns across the days of the week could be observed
SI Activity	Monitor activity through the questionnaire to identify design issues - Route taken through questionnaire - No. of error generated - Use of Help functionality - Frequency of returning to specific questions

A key organisational strategic goal is the improvement of the quality of all our statistical processes. The Eurostat report on quality (Eurostat, 2002) disgtinguishes between different types of quality. They contend that in the case of a statistical organisation the quality focus should be on the data producded and the services provided. To achieve the best quality product outcomes improving process quality is a key aim. The report explains that process quality can be imporved by identfiying key process variables, measuring these variables, adjusting the process based on these measurements and checking what happens to the product quality. This is an example of the Plan, Do, Check, Act cycle advocated by Demming (1991).

Eurostat created a handbook to identify key process variables, their measurement and measurement analysis (Aitken, et al., 2004). Key process variables are those judeged to have the largest effect on pre-defined critical product characteristics. The best indicators of quality are process variables that can be observed conveniently and continuously during the survey process and that are highly

correlated with the components of error that need to be controlled (Biemer & Lyberg, 2003). Table 3 is a list of key process variables for the data collection processes identified in the Eurostat handbook. Call History data is being collected as part of the Household Survey Transformation Project. Providing a solution for the Audit Trail data means it should also be possible to extract the key performance variables from that data source. Access to this data puts the CSO in an excellent position to monitor and improve the data collection processes based on the Eurostat defined key performance indicators. The Data Source column in Table 3 indicates the potential data source to access the variable from the CSO systems and data sources.

Table 3. Data Collection Key Process Variables

Process	Process Variable	Measurement	Data Source
Data Collection – survey instrument performance	Ability of respondents to answer a problem question	Analysis of responses and comments relating to the question	Survey data
	Number of editing errors by question	Number of errors activated	Audit Trail data – check number of errors activated by survey instance
	Route through interview	Number of deviations from automatic route	Audit Trail data – check next action after Exit from field
Data Collection – Interviewing Activity	Number of non-responses	Count of unanswered questions	Survey data
		Frequency of unanswered questions	Survey data
	Use of Help function	Count of Help activation	Audit Trail data – check number of time HELP is activated by survey instance/interviewer
		Frequency of Help by questions	Audit Trail data– check number of time HELP is activated by question
Data Collection – Interviewer Performance	Interview time	Duration of interview	Audit Trail data
	Travel time of interviewer	Duration of travel time to/from interview	Call History data
	Working hours by survey	Duration of total time including travel, admin and interviewing	Combined Call History, Admin and Audit Trail data
	Contacts by time period	Aggregate of case contacts by day and time	Call History data

Duration at all levels such as total interview time, module length and question time are considered good quality indicators that can be used to assess the quality of the survey instrument and monitor interviewer performance to assess that the questions are being asked correctly. Previously, it was not possible to access this level of detail; now this data will be easily available and can be compared across interviewers and modes. The data from the audit trails can be used to identify problem areas in the survey instruments and identify if they are design or interviewer training issues. The information can be used to target interviewer training sessions to address problem questions thus ensuring the data collection is more efficient and effective. Alternatively, insights from analysis of the data may indicate a questionnaire design issue that can be reviewed to improve the questionnaire.

The data can be used as part of a continuous improvement process, reviewing and monitoring the key performance indicators to improve the effectiveness of business processes. For example, innovation at the business process level can be achieved if the results from new questionnaire testing phases are used to inform and improve future design and development phases. The data can be used to evaluate new questionnaire designs, to identify potential problems with questions, routing or answer types. The test results can be used to improve new questionnaires ensuring the best possible questionnaires are being used for data collection, avoiding expensive data collection difficulties. The audit trail data provides the richest source of data into what is going wrong in a survey instrument and understanding the usability of questionnaires for interviewers (Olson & Parkhurst, 2013).

In addition to improving the quality of processes there are a number of known errors that can manifest during the data collection process that need to be managed (EuroStat, 2004), Table 4. Interviewers can have an effect on what is recorded in a computerised interview; in most household surveys the interviewer directly inputs respondents' answers into a computer. Variability across interviewers can lead to variation in what is recorded. In addition interviewers can affect what respondents report and respondents can affect interviewer's behaviour, for example, some interviewers may probe more and some respondents may be more likely to seek clarification (Olson & Parkhurst, 2013).

Table 4. Known Data Collection Errors

Error	Description	Measure and Corrective Action
Interviewer Error	Interviewer errors are associated with effects on respondents' answers stemming from the different ways that interviewers administer the same survey. Examples of these errors include the failure to read the question correctly (leading to response errors by the respondent), delivery of the question with an intonation that influences the respondent's choice of answer, and failure to record the respondent's answer correctly.	Analysis of question durations can be an indicator of questions not being asked as specified. If a problem is identified training session can be used to address the issue. Call recordings will be available from the call centre when the telephone interviewing starts. This data can be used to compare both interviewing modes and identify baseline durations for questions.
Measurement Error	Measurement error refers to errors in survey responses arising from the method of data collection, the respondent, or the questionnaire (or other instruments). It includes the error in a survey response as a result of respondent confusion, ignorance, carelessness, or dishonesty; the error attributable to the interviewer, perhaps as a consequence of poor or inadequate training, prior expectations regarding respondents' responses, or deliberate errors; and error attributable to the wording of the questions in the questionnaire, the order or context in which the questions are presented, and the method used to obtain the responses.	This may be due to poor survey design. Analysis of question durations that might indicate long pauses, high error counts or frequency of accessing Help can be indicators of difficulties for respondents understanding questions that can be improved with a better questionnaire design.
Non-response Error	Non-response errors, occur when the survey fails to get a response to one, or possibly all, of the questions. Non-response causes both an increase in variance, due to the decrease in the effective sample size and/or due to the use of imputation, and may cause a bias if the non-respondents and respondents differ with respect to the characteristic of interest.	Count of Don't Know and/or Refusal answers to identify problems with specific questions and/or issues with interviewer performance.

The new structure for the audit trail data allows different business users to access the data in a way that suits them and their needs, via reports or in a self-service way that allows exploration of the data. There is potential for considerable value to be realised as a result of this implementation however, it will take some time before the full impact of using the data in this way can be assessed.

7. Conclusion

This project has used a previously underused data source to solve the original business problem with the call centre contracts and provide complete transparency of CSO's household survey data collection operations. Business users have new opportunities to use the audit trail paradata to manage, monitor and improve the data collection processes. Using the audit trail data to provide insights to manage the data collection process is superior to the alternatives such as relying on subjective interviewer feedback or anecdotal data; the audit trail data is quantifiable and can be used to identify trends or problems (Duncan, 2015).

Easy access to this data provides the potential to develop a more mature approach to information management that can transform decision making at both strategic and tactical levels (Duncan &

Buytendijk, 2015). This project comes at a time when CSO is in transition and focused on improving the quality and precision of all its processes and statistical outputs. The audit trail data is an important data source that can be used to monitor new data collection process indicators and in Peter Drucker's[1] words 'What gets measured gets managed'. The data can be used to manage key process indicators that can be used to improve the quality of the data collection processes and consequently the quality and precision of the statistical outputs. Data analytics, such as this, can provide a stimulus for business areas to take transformational action steps and enable operations to be more data driven, using evidence based decision making to drive process improvements.

8. References

Aitken, A. et al., 2004. *Handbook on improving quality analysis of process variables,* Brussels: European Commission Eurostat.

Biemer, P. P. & Lyberg, L. E., 2003. *Introduction t Survey Quality,* s.l.: Wiley.

Bucher, A., 2013. *5 top trends for call centres and the pricing model dilemma.* [Online] Available at: http://www.trinityp3.com/2013/09/call-centres-pricing-model/ [Accessed 24 Feb 2016].

Corr, L. & Stagnitto, J., 2012. *Agile Data Warehouse Design.* 1st ed. Leeds: DecisionOne Press.

Devonshire, J., Liu, Y. & Cheung, G.-Q., 2012. *Blaise Audit Trail Data in a Relational Database.* s.l., Survey Research Center, University of Michigan.

Duncan, A. D., 2015. *Take an Emotional Approach to Business Analytics to Develop a Data-Driven Culture,* s.l.: Gartner.

Duncan, A. D. & Buytendijk, F., 2015. *How to Establish a Data-Driven Culture in the Digital Workplace,* s.l.: Gartner.

Eurostat, 2002. *Quality in the european statistical system - The way forward,* Luxembourg: Office for the Official Publications of the European Communities.

EuroStat, 2004. *The European Self Assessment Checklist for Survey Managers,* Luxembourg: Office for the Official Publications of the European Communities.

Nestorov, S., Abiteboul, S. & Motwani, R., 1998. *Extracting schema from semistructured data.* New York, ACM, pp. 295-306.

Olson, K. & Parkhurst, B., 2013. *Collecting Paradata for Measurement Error Evaluations.* Nebraska: University of Nebraska - Lincoln.

Snijkers, G. & Morren, M., 2010. *Improving Web and Electronic Questionnaires: The Case of Audit Trails.* Heerlen: Statistics Netherlands, Department of Methodology and Quality.

West, B. T., 2011. Paradata in Survey Research. *Survey Practice,* 4(4).

[1] Peter Drucker (1909-2005) American management consultant, professor and writer.