CATI follow-up Application

Resolving Communication links between applications using XML

Joseph Cummings & Stéphane Lamarre, Statistics Canada

Designing a Blaise application to be used in a Multi-Mode environment has resulted in a strong tool that can be used for CATI, CADE and failed edit follow-up. Having Blaise interface with external systems such as the Internet Collection System in this Multi-Mode environment has given us some interesting design challenges in terms of methods for transferring the data between the systems and evaluating the completeness of the data.

Complexities Surrounding the Blaise CATI Applications

The Regional Office Business Surveys have several complexities to them that have an impact on how we load and evaluate the data fed from the internet collection tools. These complexities involve the infrastructure used for supporting collection and the design standards for the applications.

The Collection Infrastructure

The collection infrastructure encompasses a distributed collection environment. There are several regional offices across Canada and the design of the applications make it possible for the sample to be split between these regional offices, with a copy of the applications in each regional office. The internet collection system is centralized with one core sample so there needs to be an identifier available that can split and deliver the collected data to the proper location for loading.

The collection tools interact with two external systems during collection, each of which has an impact on how the sample is fed to the internet collection system and how the collected data is loaded back into the Regional Office Business Surveys.

The Business Register: The first external system is the Business Register (BR) which is a system that contains information on all of the businesses across Canada and serves as a source for identifying the samples for the different business surveys. This information is stored in such a way that the structure of each business is easily identified and allows methodologists to draw surveys samples at different levels within the structure. The information contained in the BR includes details on the businesses, information on the respondents and information involving the collection burden on each of the respondents. The collection applications feed updates pertaining to this information to the BR nightly. This is a risk because the BR is a critical source of information related to the business across Canada. Every effort must be made to validate and verify the accuracy of these updates prior to applying them to the BR system.

The Business Survey Regional Database: The second external system is the Business Survey Regional Database (BSRD). The BSRD is a SQL Server Database with a VB.Net front end that is used to manage and maintain the samples during collection and between collection cycles. In addition to containing information about the businesses and respondents, the BSRD contains data and flags about the details on how each case is to be collected. Examples of this are the collection method (Mail, Telephone, Fax, and Internet), Birth and header verification flags (If the Birth flag or the header verification flag is on

then telephone collection is required) and case evaluation (If the evaluation of a previous collection cycle requires follow-up then telephone collection is required). The BSRD sets flags for each case that identify if the case can fit the long term and short term profile for being collected over the internet. The long term flag means that the Internet collection method can be selected if the respondent chooses this option and the cases will move to Internet collection during the next collection cycle. The short term flag is turned on or off depending on how successful the previous collection cycle was and how the birth and header verification flags are set.

Though both the BSRD and the BR system seem similar they both provide a crucial service. The BR is a secure source of information pertaining to the businesses across Canada. The BSRD is a live placeholder for the current sample information. All updates to the BR require verification. As a result these updates are not always instantaneous. The BSRD provides a method of maintaining updated business information provided by respondents until the BR system accepts or rejects the updates. Once this step is complete, the information on the BSRD is updated based upon the decision applied to the BR system.

The Blaise Design Standards

There were several requirements associated with the Regional Office Business Surveys that complicate the overall design of the applications.

Examples of this are:

- **Collection Modes in Blaise**: All of the business applications have been designed to be multimodal. The first mode of collection is Data capture and involves a data capture team that collects the data that has been mailed or faxed in. This method has a simplified flow that allows the users to skip over the intro and profiling flows and quickly access the questionnaire block. In addition, the interactive editing is suppressed but edits still trigger in the background. The second mode of collection is Telephone capture. This method is for an interviewer to collect the data with a respondent on the telephone. All of the flows are activated in this mode and are designed to establish contact with the respondent and collect the data. The edits are interactive in this mode and guide the user to correcting the data or suppressing the edits, as necessary. The third mode is Failed Edit Follow-up. This mode is for an interviewer to complete collection with a respondent on the telephone. There is enough information in the application to quickly guide the interviewers to what is needed to be done to finalize collection, whether this is due to missing information or unresolved edits. The mode Failed edit Follow-Up is necessary for resolving errors with the data loaded in from the Internet Application.

- **The Collection cycle**: The collection cycle for the Regional Office Business Surveys is monthly. The surveys are open for collection between 18-23 business days each month and closed to collection for one to two business days each month for cycle transition and sample management. There are several thousand records that require collection each cycle. This causes complications in balancing how long we can wait for a respondent to send their data though the Internet collection method before they are contacted via Telephone Capture.

- **Mixed Frequencies in Sample**: The sample within each application is a mixture of different collection Frequencies. The various frequencies are Annual, Semi-Annual, Quarterly, Every 4

months, Monthly and Every 4 weeks. These differing frequencies impact the size of the sample from one cycle to another and require a repository to manage and maintain the collection characteristics for each unit outside of the collection applications. The differing frequencies also impact the editing since the information collected is impacted by the length of time between collection periods. The frequency every 4 weeks also adds an extra complication because while the collection applications are monthly, meaning twelve periods, the frequency every 4 weeks has thirteen collection periods. One collection cycle each year will have two periods of data to collect for these units. The cycle that will contain the two periods depends on the how the End Date falls for each period. This means that the cycle for this period can be different for each unit and can change from year to year.

- **The Collection Roster**: The monthly applications are rostered. This means that there is an array of questionnaires in each application. The first element represents the current collection period and the following array elements represent previous collection periods from the most recent to the oldest period. The applications can contain up to fifteen previous collection periods. During a cycle where a case requires a thirteenth collection period, the first two elements contain the questions for the current collection periods. If a record is not fully collected for a given cycle then it is opened for collection in the following cycle. In a given collection cycle an interviewer could be required to collect data for the current period and the previous two collection periods. For most cases, this could result in three questionnaires requiring collection in a given case. For the frequency Every 4 weeks, this could result in up to four periods requiring collection in a given case. The previous collection periods usually require the respondent to be contacted over the telephone for completion. This causes a conflict in methods of collection if the current cycle is set for collection over the Internet.

- **Edit Based Evaluation**: The evaluation for each case in the Regional Office Business Surveys is derived from the edits. There are five different types of edits. Mandatory edits are edits that require correction in order to complete collection. These edit types are used when there is an obvious problem with the data collected. Severe edits are edits where the information collected is outside the norms for the case and may or may not be correct. These edits require the data to be corrected or for the edits to be suppressed and confirmed. Non-Severe edits are more akin to confirmation on the data collection and do not impact the evaluation. Incomplete Edits are edits used to identify empty values in fields that are considered critical for the completion of collection. Partial Edits are edits used to identify empty values in fields that are not critical for the completion of collection but would be good to have.

- **The Evaluation**: The evaluation is based upon two flags labelled Stage and Status.

  The Stage flag identifies where the case is in terms of collection:

  - Stage 1: Mailing is how the flag is set at initialization

  - Stage 2: Received means that the respondent has mailed back the questionnaire and it requires Data Capture

  - Stage 3: Processing means that collection has begun but the case requires follow-up with the respondent

  - Stage 4: Final means that collection for the case has been completed and it requires extraction

- Stage 5: Output means that the case has been extracted and transmitted

The status flag identifies the quality of the completion for a collected case:

- Incomplete: Collection ended and some or all of the mandatory fields have been left empty (This is set if Incomplete edits are left unresolved)
- Partial: Collection ended and some or all of the non-mandatory fields have been left empty (This is set if Partial edits are left unresolved)
- Unresolved Edits: Collection has ended but there are edits that have not been corrected or suppressed (This is set if Mandatory or Severe edits have failed and not been dealt with)
- Complete: Collection for a given case has been fully completed and is clean in terms of the edits.
- In addition to the status codes set, there are several other status codes that a user can set that relate to business. Some examples of this are Out of Business, Seasonally Closed and Out Of Cycle.

Since there can be one or more cycles requiring collection for a given case, there are two levels to the Stage and Status Flags. Each cycle has its own Stage and Status flags. These are called the Period Management Stage and Status and are derived from the edit results or a user code. At a higher level there are the Case Management Stage and Status flags and these are derived from the Period Stage and Status flags. There can be up to four Period Stage and Status flags set. The Case Management Stage flags are set based upon the worst Case Scenario period Stage and Status.

- **Header Verification**: Header Verification is completed once a year for each case though the timing of it is staggered across the different regional offices that do the collection of these surveys. Regardless of the collection method preference of the respondents, when header verification is turned on, the respondents will be contacted via telephone in order to verify the label information. The information verified covers details about the business and respondent information. The mandate is to complete Header Verification within two collection cycles so a respondent could be contacted up to two collection cycles each year via telephone despite the preferred collection method.

The Design Solutions for Loading the Internet Data

In order to load the data from Internet collection into the Regional Office Business Surveys, solutions were put into place to address the design complexities in the Blaise applications: Implement two-way communication between the Internet Collection System and the Regional Office Business Surveys; Build a placeholder for the label updates; Add controls to manage the collection responsibility and add extra controls to the application to aid in loading and evaluating of the data from the Internet Collection System.

- **Two Way Communications**: In order to ensure that the Internet Collection system has the information necessary to send the collected data back to the appropriate collection site, it was thought

best that the sample sent to the Internet Collection System should come from the Regional Office Business Applications. This serves three purposes. First, because the sample is loaded in the CATI application and, then, extracted for the Internet Collection System, the Blaise application is already set up and ready to receive any respondent data returned from the Internet Collection System. Secondly, The Blaise CATI application serves as a central point for sorting through all of the criteria involved in choosing which cases can be sent for Internet Collection. Thirdly, by sending the sample from the Blaise application, we are able to send the Regional Office location for each record so that the Internet System will be able to return the respondent data back to the correct location. The second purpose for the two way communication is to update the Internet collection system with any information related to the collection cases through other modes of collection. This way records collected though other modes can be removed from the Internet Collection System before burdening the respondent with E-Mail reminders.

- **Place Holder for Business & Contact Information Updates**: In order to prevent updates from being directly sent to the Business Register prior to verification, we have built a place holder inside the Blaise applications for this information. All information related to updates to the business or the contact information is loaded into the fields in the place holder. When a user accesses a case for telephone follow-up, the user will have access to these updates through a parallel tab. Once the information is confirmed, the user can copy the information from the place holder tab and paste it in the relevant fields where it will ultimately be updated in the Business Register. The Place Holder tab also has an impact on the evaluation. If the case is fully collected and clean but there is information in the Place Holder tab then the evaluation stage will stay Processing and the evaluation Status will become Review. There is a Checkbox called reviewed in the Place Holder Tab. Once all of the business & contact information has been reviewed, the user can check this box to prevent the Place Holder Tab from having an impact on the evaluation.

- **Controls for Collection Responsibility**: The four Methods of Collection that a case can be in at the beginning of collection are Mail, Telephone, Fax and Internet. The full sample for each collection site is loaded into the Blaise Regional Office Business Surveys so controls need to be added to prevent the Mail, Fax and Internet cases from being added to the Blaise Scheduler. Within the application, each record has a Mail_Internet_Release flag. For all cases slated for telephone capture this flag is set to a 1, for all other cases, this flag is left empty. Cases will only be added to the scheduler for telephone capture if this flag is a 1. Since the collection cycle is so short, there is a release date set for each cycle that will release the Mail and Internet cases for telephone capture half way through the collection cycle in order to ensure every effort is made to collect the necessary data. The release controls simply set the Mail_Internet_Release flag to 1 in order to redirect them for telephone capture. The flag can also be set back to empty for specific cases where the respondent has already made efforts to deliver the data through one of the other collection methods.

- **Controls for Loading and Evaluating the Data**: Due to the complexity of the collection applications, extra controls needed to be added throughout the flow to ensure a proper evaluation is set when loading the data and running the Check Rules in both the DEP and through Manipula. The first issue involves the Access level. There are two Access Levels used in the application that affect the evaluation (Interviewer and Senior Interviewer). Senior Interviewers, by description, have more experience and responsibility than Interviewers. When a case is evaluated, certain status codes will

automatically finalize the case for a Senior Interviewer and stay processing for an Interviewer. When loading the data from the Internet system, we set the Access Level to Interviewer in order to ensure the Stage is not set to Final prematurely. The second issue relates to flows and triggers in the application. Auxfields are used to trigger occurrences such as the evaluation within the application. When running the Check Rules, the auxfields are empty so certain components we want to trigger stay inactive when the check rules is run. To solve this we have added a field called BatchMode to the applications. This field remains empty throughout collection with the exception of when we load data from the Internet System via Manipula. At this point the BatchMode flag is set to one, the data is autocopied, the Check Rules is run and, finally, the BatchMode flag is reset to empty. Throughout the application, we add the condition for the BatchMode being set to one along with the auxfield condition. This ensures the rules trigger properly during collection in the DEP and during the loading of the data from the Internet system via Manipula.

Receiving Data from the Internet System

There have been two different systems designed for collecting data over the internet. The first system developed was called Electronic Data Reporting (EDR). The second system is called E-Questionnaire. Neither of these two systems uses Blaise technology due to strict security requirements within Statistics Canada.

- **The Electronic Data Reporting system**: The Electronic Data Reporting system was developed in parallel with the Regional Office Business Surveys. This system was complicated and burdensome to interface with. How the communication was achieved was through a process of field mapping. During the development of a questionnaire in the Electronic Data Reporting system, a key file would be provided from the developers of the Blaise application. The key file contained a list of the names of all of the blocks and fields that were developed in the Blaise application.

  An example of this is as follows:

```
MSM
Profile.Legal_Name
Profile.Operating_Name
Profile.Contact_Name
Comments.EDR_1
Comments.EDR_2
CurrentMonth.Questions.GOM
CurrentMonth.Questions.Total_Shipments
```

  This key file was then loaded into the EDR system where it was linked to the matching fields in the EDR system. Each night during the collection cycle, the data for all cases that were collected in the EDR system were extracted into parsed files and sent for loading into the Blaise applications. A generic tool was designed to load the data, regardless of the design in Blaise so a generic format was required for the files containing the data extracted from the Electronic Data Reporting System. The file was semicolon delimited and consisted of three fields: the Primary Key, the Blaise Field, the value to be loaded.

An example of one of these files is as follows:

```
232345345;Profile.Legal_Name; My Technologies
232345345;Comments.EDR_1; Company Closes over the winter
232345345;CurrentMonth.Questions.GOM;3452354.23
232345345;CurrentMonth.Questions.Total_Shipments;3452354.23
12341234;Profile.Legal_Name; Your Technologies
12341234;Comments.EDR_1; Expanded locations by 2 this year
12341234;CurrentMonth.Questions.GOM;125.23
12341234;CurrentMonth.Questions.Total_Shipments;345.66
```

This process was found to be very time consuming. The problem grew in scope and scale with the size of the sample and the number of questions to be loaded. This solution also contained several points of failure. With a record in the file for every field in a given case, it was very difficult to validate that all of the information successfully passed from one system to another. If information was missing then there were two possible points of failure: The extraction process from the internet system or the loading process into the Blaise application. Time and effort was spent daily to ensure nothing was lost in this process.

- The E-Questionnaire system: The E-Questionnaire system is a more recent system developed at Statistics Canada. This system was designed to make use of XML as its core way of communicating with other applications. When first interfacing with the E-Questionnaire system our Blaise applications were not ready to communicate with XML. In order to resolve this, a tool was designed to convert the XML files into flat files that could be loaded into the Blaise applications. This is a significant improvement over the previous method of loading data from the EDR system since there is one record per period in a given case. Currently only the current period is collected through E-Questionnaire but this design allows this to be expanded to collect previous periods if it is deemed necessary.

  The flat file is delimited with the ~ character. The layout is designed to account for the common fields first, then move to the survey specific fields towards the end. The common fields consist of the Primary Key, the Place Holder fields, the E-Collection Comments, the Reference Period and the survey specific fields.

  An example of one of these files is as follows:

```
232345345~LName Change~~~~~Comment 1~Comment 2~~~~~201006~1234.45~~356~45~
12341234~~~~New Contact Name~Contact Phone~Comment 1~201006~112~3456~356~4~
```

  It is not a perfect system though since there is an extra point of failure when converting the file format from XML to a flat file. This process requires a validation step to ensure errors are not introduced into the data.

  Currently at Statistics Canada, we are reviewing Blaise 4.8 and the improvements to performance that it offers. One of these improvements is the ability to use the file format of XML for sending and receiving data. We have successfully managed to introduce Manipula code from Blaise 4.8 using the Manipula.exe from 4.8 into Blaise 4.7 applications. The experimenting done to this point shows great promise in communicating with external systems such as E-Questionnaire. As part of our experiments, we have created basic XML sample files and have tried different methods of loading this data into the application.

An example of one of these files is as follows:

```
 <?xml version="1.0" encoding="UTF-8"?>
<Dataset><Datarecord>
<PriKey>01000001</PriKey>
<Com1>1st_Comment</Com1><Com2>Middle_Comment</Com2>
<Val1>201006</Val1><Val2>1234.45</Val2><Val3>356</Val3>
</Datarecord>><Datarecord>
<PriKey>02000001</PriKey>
<Com1>Another_Comment</Com1><Com2>Comment_Again</Com2>
<Val1>112</Val1><Val2>3456</Val2><Val3>4</Val3>
</Datarecord></Dataset>
```

The code is fairly simple to implement but powerful enough to introduce sophisticated requirements when necessary.

A simple example of this code is as follows:

```
SETTINGS
   DESCRIPTION = 'XML to BLAISE'
USES
   DATAMODEL xmlModel
       FIELDS
            Prik : STRING[9]
            Com1 : STRING
            Com2 : STRING
            Com3 : STRING
   ENDMODEL
   OutputMeta '..\MSM'

INPUTFILE InputFile1: xmlModel ('Trialxml.xml', XML)
UPDATEFILE UpdateFile1: OutputMeta ('..\MSM', BLAISE)

MANIPULATE
  UpdateFile1.get(InputFile1.Prik)
  IF UpdateFile1.RESULTOK THEN
       UpdateFile1.comments.ro1 := InputFile1.com1
       UpdateFile1.comments.ro2 := InputFile1.com2
       UpdateFile1.comments.ro3 := InputFile1.com3
       UpdateFile1.CM.CM_BatchMode := 1
       UpdateFile1.checkrules
       UpdateFile1.CM.CM_BatchMode := EMPTY
       UpdateFile1.WRITE
  ENDIF
```

The major improvement we have found in moving to XML is in reviewing the sample file provided. With the expandable and collapsible tags, it is simple to verify the quality of the data and identify any errors that may be contained within.

We are only in the beginning stages of converting our applications to make use of Blaise 4.8. Converting our file formats to XML will be one of the critical steps in this conversion as we are trying to make XML the standard format for communicating data between the various systems we have at Statistics Canada.