Data Extraction, A Hindrance to Using CAATs

Tommie W. Singleton, Ph.D., CISA, CGEIT, CITP, CMA, CPA, is an associate professor of information systems (IS) at the University of Alabama at Birmingham (USA), a Marshall IS Scholar and a director of the Forensic Accounting Program. Prior to obtaining his doctorate in accountancy from the University of Mississippi (USA) in 1995, Singleton was president of a small, value-added dealer of accounting IS using microcomputers. Singleton is also a scholar-in-residence for IT audit and forensic accounting at Carr Riggs Ingram, a large regional public accounting firm in the southeastern US. In 1999, the Alabama Society of CPAs awarded Singleton the 1998-1999 Innovative User of Technology Award. Singleton is the ISACA academic advocate at the University of Alabama at Birmingham. His articles on fraud, IT/IS, IT auditing and IT governance have appeared in numerous publications, including the *ISACA Journal.*

Almost all auditors agree that a key tool in conducting audits, especially fraud and IT audits, is the use of a computer-assisted audit tool (CAAT). There are many factors that go into the effective and efficient use of CAATs in IT audits, including technology issues, social/personnel issues, choosing the right CAAT, defining the data to extract and making sure audit objectives drive the use (or fit) of a CAAT.

Anecdotal evidence suggests that one of the primary hindrances, if not *the* prime one, of using CAATs is in getting the data from the operational system into the IT auditor's CAAT. This article will center on data extraction, focusing on the most efficient methods given the current state of features among the leading CAATs vendors.

IDEAL IMPORT FORMAT

The ideal format of data being imported into a CAAT is generally a flat file in which the first row contains the column headings and the second row begins the data set and in which the data set (rows) is contiguous until the end of the data (see **figure 1**). That is, subtotals, breaks and subheadings create situations where data have to be "cleaned" or manually manipulated into the ideal format. This format is the goal of data extraction, regardless of the specific methodology.

DATA EXTRACTION DATA FORMATS

The IT auditor will need to consider the different formats of data available for data extraction and find the best fit for the tool and operational data format. Factors that affect this decision are platform, database/database management system (DBMS) and application software (i.e., the accounting software system).

The data extraction file could be one of several formats, such as dBase, PDF, Excel, Extensible Markup Language (XML), delimited text and open database connectivity (ODBC), to the operational data files. Some of these are easier or more efficient for extraction purposes. Generally speaking, the order of ease with which to work follows the order in **figure 2**.

Caution should be used in converting operational data into some of these formats. For example, when converting data into a PDF file, it is important to make sure that the file is not a scanned image (which will not work). Usually, printing to a PDF file is easier to work with than saving the data as a PDF. Most "heavy duty" CAATs today can read data from a PDF file, even if it is a report filled with breaks, subtotals and extraneous data—in other words, a report in which the data get messy. The CAAT features allow the IT auditor to pick and choose the data with relative ease from the PDF soft-copy document.

When exporting to a text file (ASCII format), systems often add breaks, subtotals or subheadings. The text file should follow the "ideal" format demonstrated in **figure 1**. Also, the fields (columns) should be delimited with a comma (CSV) or tab; for the data to read correctly, a delimiter is usually necessary.

Figure 1—Ideal Data Format for Data Extraction

ID	NAME	ADDRESS	CITY	STATE	ZIP	PHONE	CONTACT	CREDIT LIMIT	BALANCE
1	ABC Co.	101 Main St.	Anywhere	FL	33333	123-4567	Joe	$50,000.00	$24,000.00
2	Cranky Repairs	211 Elm St.	Anywhere	FL	33333	234-5678	Sally	$10,000.00	$1,200.00
3	Mild Soap Inc	314 Oak Ave.	Anywhere	FL	33333	345-6789	Tim	$30,000.00	$5,000.00
4	Sunny Side Home	411 Pine St.	Anywhere	Fl	33333	456-7890	Sue	$25,000.00	$26,000.00

FORUM

Do you have something to say about this article?

Visit the *Journal* pages of the ISACA web site *(www.isaca.org/journal)*, find the article, and choose the Comments tab to share your thoughts.

Figure 2—Data File Formats by Order of Efficiency to Import

File Format	Type/Use
dBase	.dbf
Adobe PDF file (not a scanned image)	Export data or print data to a PDF file.
Microsoft Excel file	Export data as an .xls file.
Delimited text file (e.g., CSV)	Export or save as an ASCII/text file, with delimited fields (comma or tab).
XML (XBRL is similar to XML.)	.xml
Others (e.g., bring data file over directly into CAAT)	The others are fairly time-consuming to use.

OPTIONS TO EXTRACT DATA FROM OPERATIONAL SYSTEMS

There are usually one or more ways that a platform/system will allow the IT auditor to pull data from the operational system to extract the data needed. These options will be discussed beginning with the one that is generally considered the most efficient method.

First, one should investigate the export functionality options of the accounting application. Some usual options are "save as" options that include Excel, PDF or text delimited files. If the IT auditor can load a report or data file that contains some or all of the data needed, a save-as option may be available, especially in Microsoft-type systems. It could also be a menu option that allows data to be extracted (e.g., MENU -> FILE -> EXPORT). This option is usually the easiest one to perform, and it can usually export data into the easiest-to-use formats (see **figure 2**). The save-as function can serve as an export function as well.

Sometimes the best approach is to extract the data in one format and then convert it to a PDF file. For instance, a "messy" Excel spreadsheet can be efficiently cleaned up by converting it to PDF (i.e., print to PDF), and then using the CAAT to identify and extract the data from the report.

Second, if necessary, one should investigate the print and report functionalities. For example, many systems allow reports to be printed as a soft-copy file, rather than a hard-copy printout. In the print dialog box, this option would be available if the system allows for "print to file." Print to file creates a text file output of the report. It is important to note that there may be a need to convert the text file into the ideal format (see **figure 1**). A better option is to print to PDF. Many systems have that option, even if Adobe Acrobat is not installed. If the system allows the data file or report to be printed to a PDF soft-copy file, it is important to note that this method is the second-easiest file format (see **figure 2**). Of course, the IT auditor could simply print the data needed to a hard copy and manually key it in to the CAAT, but this option should be used as a last resort as it is time-consuming.

Last, if needed, one should pull the data directly from the operational database into the CAAT. This can be done with ODBC, a dynamic connection from the CAAT to the operational database. It is usually possible to extract the data using Structure Query Language (SQL), because SQL is used by almost all databases. Additionally, XML is becoming a common data extraction and communication tool. Microsoft products and many accounting applications are compatible with XML. But this option requires a few things the others may not require. The IT auditor will probably need a data dictionary to extract data using ODBC, SQL or XML, and the data dictionary may not be readily available.

DATA INTEGRITY

Before using the extracted data in the CAAT, the IT auditor will need some assurance that the data set in the CAAT is identical to the data on the operational system. There are various ways of performing a "crosswalk" or reconciliation, but the IT auditor must make sure to select some reasonable method to ascertain integrity of the CAAT data. Often, this involves something similar to the old batch transmittal sheet methodology. In that methodology, one created metrics about the data set, e.g., number of records, total dollar amount column, total numeric column and other similarly identifiable summary facts.

CONCLUSION

CAATs provide a method for IT auditors that is efficient and effective in meeting audit objectives. In fact, IT audit pioneers stated that the invention of CAATs was the most significant event in the history of IT audit. But that does not necessarily mean that it is always easy to use a CAAT. Perhaps the most difficult step in using a CAAT is getting the data in a usable format in a reasonable amount of time. The information in this article is intended to make that process as efficient as possible with any given platform, database and accounting application.