Data Portability: It is about the Data—the Quality of the Data

ABSTRACT------------------------------

The life cycle of software applications used to capture and manage data is but a fraction of the life cycle of the data itself. The issues of data portability and long-term data preservation are now critical as companies are realizing that they must be proactive in protecting their data. Understanding how standards influence data portability is an important first step in managing intellectual assets and avoiding software lock-in.

BIOGRAPHY------------------------------

Peter R. Benson
Executive Director and Chief Technical Officer
Electronic Commerce Code Management Association

Peter R. Benson is the Executive Director and Chief Technical Officer of ECCMA; he is an expert in distributed data systems and master data management. Peter developed data collection and reporting systems for large agricultural businesses as well as for the public relations, advertising, healthcare and financial industries. Peter was granted a British patent in 1992 covering the maintenance and usage tracking of distributed data. Peter developed the UNSPSC, an internationally recognized commodity classification for spend analysis and went on to develop the eOTD a leading open technical dictionary used to create unambiguous language independent descriptions of individuals, organizations, locations, goods, services, processes, rules and regulations. Peter served as the elected chair of ASC X12E the US Standards Committee responsible for the development and maintenance of EDI standard for product data. Peter is part of the US expert delegation to ISO TC 184/SC 4 and the project leader for the international standard ISO 8000 (data quality) and ISO 22745 (open technical dictionaries). Peter has a baccalaureate in mathematics and physics from the Academy of Bordeaux, France, a bachelor of science in agriculture from London University, England and a master of science in agricultural marketing from London University, England.

Data Portability, the antidote to data "lock-in"

It is about the data - the quality of the data

Mr. Peter R. Benson

Abstract:

The life cycle of software applications used to capture and manage data is but a fraction of the life cycle of the data itself. The issues of data portability and long-term data preservation are now critical as companies are realizing that they must be proactive in protecting their data. Understanding how standards influence data portability is an important first step in managing intellectual assets and avoiding software lock-in.

Background:

There was a time in the distant past when you bought a "Computer" and it came as a complete and inseparable package, there was no concept of hardware or software. Then came the operating systems UNIX, then the Disc Operating Systems, or DOS with Microsoft DOS (MSDOS) and Apple DOS, a few may even remember CPM and MPM. This was revolutionary in that operating systems allowed the separation of the hardware from the software; competition and innovation followed.

As operating systems evolved we got plug and play where hardware and software talked to each other even to the point where sometimes it was hard to interrupt the conversation. As the years went by we started experiencing problems of backwards compatibility, the ability to access old data with the new versions of a software application. Worse some software vendors went out of business, even big ones (remember MicroPro and WordStar?) and remember Y2K? Upgrading COBOL was really an exercise in moving data and it became an expensive priority. Luckily software applications were relatively simple and data was even simpler. Today this is no longer the case, hardware, operating systems, software application and data are much more complex.

In our modern environment the life cycle of software applications used to capture and manage data is but a fraction of the life cycle of the data itself, the issues of data portability and long-term data preservation are critical. Companies are realizing that they must be proactive in protecting their data and manage risk. Even when data is stored locally and safely backed up, it may not be separable from the software application. In the new Software as a Service (SaaS) or cloud computing environment where data is stored remotely by the service provider, it may not be separable from the service.

A simple example of this problem can be seen in personal on-line banking, a great service but it is only when you decide to change banks that you realize that you have to re-enter all the information you need to pay your bills. If you thought this was no more than a technical oversight of the banking community you would be underestimating the time and effort they put into keeping their customers. Customer "lock-in" is a very common commercial practice and frequently an integral part of pricing strategy. Standards are the antidote to lock-in and this applies to data just as it does to plugs and sockets.

Solving the data lock-in problem requires simple awareness of what it takes to create portable data and requesting or requiring that application and service providers adopt international standards for data quality. ISO 8000 Quality data is portable data.

Understanding data

While a datum is defined as a disruption in a continuum, a more practical definition of electronic data is "symbolic representation of something that depends, in part, on its metadata for its meaning". It follows therefore that the quality of the metadata must play an important part in determining data quality. Metadata gives data meaning. For example "50-02-01" is a meaningless string of characters but apply the metadata "Date of Birth" and it becomes meaningful data. To make it unambiguous we need to have syntax such as CCYY-MM-DD and the associated value becomes 1950-02-01.

Good quality metadata comes from a metadata registry or a technical dictionary. This will contain a definition of the concept. For example, the concept: "Date of birth" has a concept definition of: "Year, month and day in which a person or an animal is born". Even better, an open technical dictionary will assign a language independent public domain concept identifier, as for example 0161-1#02-065175#1 in the Electronic Commerce Code Management Association (ECCMA) open technical dictionary (eOTD). This allows the data 0161-1#02-065175#1:1950-02-01 to be rendered as either Date of Birth: February 2, 1950 or Date de naissance: 2 Février 1950.

Using quality metadata from an open technical dictionary creates not only quality data in the sense that it is unambiguous, but it also creates portable data, data that can be easily moved from one application to another and preserved over time independently of software. Finally using pubic domain concept identifiers as metadata protects the intellectual property in the data from claims of "joint work".

Creation of a "Joint Work"

When an operator interacts with a computer they do so through a software application. As data is keyed into the computer, the software application stores it in electronic form. Close examination of the stored data reveals that it includes not only the characters keyed in by the operator but "hidden characters" inserted by the software application and also the data is organized according to a structure controlled by the software application. In essence the data stored in the file (a fixed form in copyright terms) was created through a collaborative effort of the operator that keyed in the data and the owner of the software application; it is a "joint work" where both parties contributed intellectual effort to its creation. The authors of a joint work are co-owners of the copyright in the work, unless there is an agreement to the contrary. Removing or replacing the proprietary metadata or changing the structure of the data is a violation of copyright unless it is done with the permission of the "joint owners". If the software application includes a feature to export the data in a neutral or portable format then the resulting data is no longer a joint work.

Claims that an application exports data in XML does address the syntax part of the problem, but that is the easy part. What is required is to be able to export all of the data in a form that can be easily uploaded into another application. This can be a time consuming and expensive exercise not to be undertaken lightly, a software upgrade will always be cheaper, this is the hallmark of software lock-in.

Implementing international standards for data portability

While ISO 10303 is the standard that defines the neutral exchange format for design, manufacturing or production data, ISO 8000 is a more general standard concerned with the principles of data quality, the characteristics of data that determine its quality, and the processes to ensure data quality.

ISO 10303 provides a representation of product information along with the necessary mechanisms and definitions to enable product data to be exchanged between applications as neutral, portable data. ISO 10303 addresses specific engineering environments such as automotive, shipbuilding, aerospace, defense, general manufacturing and plant management.

ISO 8000 address more general data categories. The first parts of ISO 8000 to be published deal specifically with master data, the data that identifies and describes individuals, organizations, locations, goods, services, processes, rules and regulations, the essential and fundamental data of any business.

ISO 8000-110:2008 is the foundation standard for master data quality. Master data that is compliant with the standard is portable data that is formatted according to a published syntax and where the metadata is explicit, either included with the data or by reference to an open technical dictionary.

ISO 8000-120:2009 is a supplement to ISO 8000-110:2008 that covers master data provenance. The standard is designed to assist in tracking the extraction of data elements through to their original source. Implementation of this standard requires knowledge of database management.

ISO 8000-130:2009 is a supplement to ISO 8000-120:2008 that covers master data accuracy. The standard is designed to assist in tracking claims to accuracy of data elements. Implementation of this standard requires knowledge of database management.

Conclusion

Requesting or requiring that master data is provided in ISO 8000-110:2008 compliant format is not a burden to the data provider. The requirements of ISO 8000-110:2008 are simple; they require no specialized technology or the purchase of any product or service and are within the capability of all companies regardless of their size.

In a world rapidly moving towards SaaS and cloud computing, it really pays to pause and consider not just the physical security of your data but its portability.

ECCMA is the project leader for ISO 8000. ECCMA has developed and manages a certification program and a register of ISO 8000-110:2008 certified individuals, organizations, software applications and data services; more information can be found at www.eccma.org.