Capitalizing on Big Data: Governing information with automated metadata

Stephen Turner
Known-Quantity.com, part of Turner & Associates, Inc.

ABSTRACT

Equity markets recognize the inherent value of information and emerging opportunities associated with Big Data. Hackers are even more enthusiastic about information, demonstrated by exploits targeting intellectual property and customer data. Yet, many companies do not recognize the value of their own information. In part, this can be attributed to the reality that most chief information officers (CIOs) are not tasked with valuation of intangible information assets. As a result, strategic investments in information systems are not based on protecting and growing the underlying value of information.

This paper introduces a new technological system to support the valuation of information assets. CIOs can deploy this model with the finance team and general counsel. The three building blocks of this model are metadata, automation, and learning machines. Metadata can be updated to frequently revalue each piece of datum. Automation can be scaled to process vast quantities of data based on predefined protocols and standards. Learning machines can evolve as authorized staff correct errors.

This model is designed to support corporate and institutional Information Governance (IG). Although IG is playing an increasingly important rule as Big Data emerges, there is a gap. The metadata-based system closes the gap by enabling real-time queries of assets across the enterprise. Current opinions by most state bar associations advocate for metadata to be confidential and privileged. This is beneficial in an age in which intellectual property – in the form of intangible information assets – needs to be guarded with particular care.

Keywords: Big Data, information valuation, data science, predictive coding, learning machines, Information Governance

INTRODUCTION

Big Data consists of a growing torrent of bits and bytes ranging from email messages and social media content to video surveillance and data from sensors. Corporate and institutional information systems are growing to accommodate this flood of data – but not fast enough. "Data warehouses were not designed for the volume of integration and access required," noted Lori MacVittie (2012). She added, "the sheer volume of incoming data can be at times enough to overwhelm the supporting systems" (MacVittie, 2012).

Knowledge workers commonly create and make several rounds of edits to Word files, Excel spreadsheets, Powerpoint presentations and send emails from the time they wake up to the moment they fall asleep. On average, it would take 16 hard drives for each person to manage the data users create, edit, or alter in some way each year (James, 2010). It isn't practical for people to keep all their data on their hard drive. Most users need a *secondary* form of storage – often with a collection of devices to retain their data. These include optical storage devices, such as CD or DVDs. Others have a plethora of hard disk drives. Secondary storage can also include flash memory and zip drives (James, 2010). At the *tertiary* level, the processing unit can access the data using direct attached storage (DAS) or on a network. It needs to use a file system to identify where it is stored (Logix4U, 2012). For enterprises, network-based storage has become an essential part of their data storage strategy, but the challenge is to assure that they access their data rapidly, while keeping costs down. Within network storage systems the most distant connection is called the "edge" (Logix4U, 2012). Aberdeen suggests that the way data is stored (and secured) should correspond to its value to the enterprise (Csaplar, 2012). The most valuable data can then be mainstreamed into various uses that add value.

Internal processes and prioritization methods are evolving to prioritize investment in the most valuable information and delete other data that does not have future economic benefits. With so much information being generated on a daily basis, decision-making can no longer be made on a case-by-case basis. Policies and workflows must define protocol. Tagging and valuation of data can assist in triage of where it should be stored, if at all. This would enable organizations to move data not needed for daily access to the "edge." Data with no value and no duty to maintain custody could also be tagged for deletion – saving resources.

"Information Governance [is] the specification of decision rights and an accountability framework to ensure appropriate behavior in the valuation, creation, storage, use, archiving, and deletion of information." According to Gartner, "it includes the processes, roles and policies, standards, and metrics that ensure the effective and efficient use of information in enabling an organization to achieve its goals" (2012).

Ted Friedman, a Gartner vice president, stated, "information governance is a priority of IT and business leaders as a result of various pressures, including regulatory compliance mandates and the urgent need for improved decision-making." In 2012, Gartner highlighted the importance of master data management (MDM) as part of IG. Gartner characterized MDM as a "technology-enabled business discipline" that is only possible when businesses and IT organizations collaborate on data assets. The internal goal is accuracy, semantic consistency, and accountability of those assets (Gartner, 2012).

Automated information systems can execute these priorities and humans can monitor the accuracy and reliability of these systems. The primary components of the system are metadata, automation and machine learning. All technologies described in the paper are currently in use or have been validated by other researchers. Existing technologies require integration and

cooperation of senior management. This is only possible if the parties recognize the underlying value of information. Organizations with formal policies and procedures are well positioned to realize the value of Big Data.

TECHNOLOGICAL ENABLERS OF A NEW MODEL

IG policies are essential, but MDM must expand to include real-time assessments of value and risk associated with each piece of datum. Organizations with this capability will have a significant competitive advantage in the emerging digital culture. Based on Hulten's and Hao's calculation, intangible capital contributes 44 percent of a publicly-traded company's market capitalization (Gartner, 2012). Information is intangible capital.

In today's fast paced environment, it is impractical for staff to assign these valuations by hand, which would be less consistent than machines. In order to annotate the value of data, existing technologies must be utilized across the enterprise in new ways.

Big Data tools such as Platfora (used on Hadoop) can run queries across distributed data sets. Policy annotations and valuation can be automatically assessed on a daily or hourly basis or valued each time the data is accessed or revised. The data can be appended to describe where the data came from, how it has been improved upon and types of models that might apply – such as rate of decay. The author of this paper has co-authored another paper entitled "Market Value Impact of Information Assets," which defines seven dimensions of valuation. The diagram titled Figure 1, as indicated in Appendix A, illustrates how this valuation system can be implemented.

In the absence of complete metadata at the point when information enters the organization, automated tagging will be applied to each piece of datum. Tags must comply with taxonomic standards. Data scientists may conduct quality assessments through tests or audits. They are likely to find that automated decisions are not always accurate. In addition, during the course of normal job tasks, authorized users will conduct data queries. During review of results, they may identify inaccurate metadata as well. Authorized users and data scientists would be expected to report errors to a Metadata Curator. This administrator may edit Company's Taxonomy Standard if its guidance leads to unforeseen errors. This process improves the quality of the metadata over time, especially with data accessed more frequently.

Automation has its limits, so this model assumes that users may report errors to an administrator, who will determine whether refinements are required. Refinements would start with the company's taxonomic standards, which define terminology and hierarchy of information. As business changes, new terms are introduced and new priorities are agreed upon. These refinements can be absorbed by a system that utilizes predictive coding. Machine learning enables continuous improvement to the automation of meta tagging.

Even with best practices in place, audits should be conducted to assure policy compliance. CIOs and CFOs have a responsibility to recognize these liabilities, quantify them and mitigate the potential risks. Metrics could be applied to the data of if its value increases or is depleted in the event of a breach. Large volumes of information on or within the network could be queried and evaluated to determine where there might be duplicates. Corrective action could then bring it all together.

Building Block One: Metadata

Metadata can contribute to the security of data in storage as well as when it is being

transported or transferred. Mattson asserts that metadata should be included with protected sensitive data with required information for decryption. He has demonstrated that this also works with credit card data. He calls his approach "Continuously Protected Computing" (Mattsson, 2012).

Metadata is "data about data" (Media/Outreach, 2012). The term "meta" is a Greek word that means "after, behind" or "higher, beyond" (Online Etymology Dictionary, 2012). The word "tag" was first recorded in 1835 to represent "label." Tag was first equated (in writing) with automobile license plates in 1935 (Online Etymology Dictionary, 2012).

Offline meta information has been used to label, describe, or identify objects for generations. In libraries, prior to the use of computers, a card catalogue included metadata about books, such as publisher, author and page count. Melville Dewey created the Dewey Decimal System of Classification in 1872 to assign metadatum in the form of a number (OCLC, 2012). For example, the 600 call number references Technology (applied sciences), while 300 references the Social Sciences. These numbers in the card catalogue correspond to the label or tag on a book with the same number (University of Illinois, 2012).

The International Press Telecommunications Council (IPTC) was formed in 1965 to author vocabularies for sharing of news data. By 1979, it had established metadata standards for news photos. President Reagan's photographer Mike Evans attempted to automate this process through Adobe Photoshop. Thereafter, IPTC introduced the Information Interchange Model (IIM), which evolved until 1990 when it included "schema" (Photo Meta Data, 2012). Today, digital/mobile cameras automatically embed information with each photo. By 2001, Adobe introduced the Extensible Metadata Platform (XMP) and uses Extensible Markup Language (XML) – bringing photos into the mainstream of data management (Photo Meta Data, 2012).

The United States Geological Survey (USGS) uses metadata extensively to describe a range of unstructured geospatial information. To be consistent, they comply with the Content Standard for Digital Geospatial Metadata, which specifies 334 different elements, such as Altitude Datum Name, Bearing Resolution, Calendar Date, Depth Encoding Method, Landsat Number and Purpose. One hundred nineteen of these, including Series Information, Format Information Content, and Attribute Value Accuracy, exist only to contain other elements (United States Geological Survey, 2012). This is a standard taxonomy that makes sharing possible. As an example, the specification provides the following structure to define data quality:

Data_Quality_Information =
0{Attribute_Accuracy}1 +
Logical_Consistency_Report +
Completeness_Report +
0{Positional_Accuracy}1 +
Lineage +
(Cloud_Cover) [Federal Geographic Data Committee, 2012]

Apart from USGS and IPTC, there is much photographic and video content posted online or stored in databases that are not tagged. Matt Richardson recently invented the Descriptive Camera, which processes an image in about six minutes and textually describes what it sees. His vision: "Imagine if descriptive metadata about each photo could be appended to the image on the fly" (Daily Mail Reporter, 2012).

Meta tagging is commonly used as part of the web today. Around 1995, the World Wide Web Consortium (W3C) expanded HTML to include meta (Lastowka, 1999). This information about the content of a webpage can then aide search engines in understanding whether a search for something not in the actual content may be relevant. For example, if a web user searches for a phrase that includes the word "geology," an article about a certain type of rock may not include the word "geology." In theory, the meta tag can rectify this shortcoming through keywords and descriptive meta tags. However, Google claims that it does not actually recognize certain meta tags in its algorithms any longer (Google Webmaster Central Blog, 2009).

Much of today's metadata on the web are semantic – which are contextually relevant. The other type of metadata is syntactic. The later describes its organization and what it looks like (Wolfe et al., n.d.). Meta tags about value tend to be more semantic, rather than syntactic. Although most meta tagging has been manually attributed to data, it is possible to automate the process. Products such as Magento Connect, Joomla Tag Meta Manager and MetaGenerator have been created for use on the web.

What is on the web is obviously public, but much of the metadata that this paper describes is intended to be private and accessed by authorized staff. Bar associations in New York City, New Hampshire, Maine, Florida, Arizona, and Alabama have analyzed whether metadata is off-limits – ethically speaking – during litigation discovery. They concluded that it is unethical to data mine or even review metadata, unless a sender provides consent (Perlman, 2009). The first bar association to weigh-in on the matter was in New York State. Its Committee on Professional Ethics was the stated that "a lawyer may not make use of computer software applications to surreptitiously 'get behind' visible documents." This builds upon another broader opinion by the American Bar Association, Op. 92-368, which looked at a range of "inadvertent disclosures" (New York State Bar Association, 2001). As a result, metadata is considered to be confidential by several bar associations. Associations in the District of Columbia and Pennsylvania disagree, and West Virginia is on the fence (Perlman, 2009).

Andrew Perlman disagrees. The professor of law at Suffolk University has reviewed court opinions and legal scholarship on the matter and determined that a flat ban on metadata mining by third parties who gain access inadvertently – without the knowledge of the sender – is misguided. His argument is that some metadata is confidential and some is not (Perlman, 2009). Therefore, data scientists should consider developing policies that define what is and what is not confidential or privileged. Metadata can include a notation that establishes whether a piece of datum is confidential or privileged, consistent with a predefined set of information governance policies and taxonomic standards.

Building Block Two: Automation

Relational databases already operate with automated meta tagging. Researchers Wolfe, Sanchez, and Chaple advocate for the development of standardized taxonomies by ontologists – much as the USGS uses – that can be consistently applied to information. Taxonomy Manager is a product the researchers developed and tested at the Naval Postgraduate School. In 2010, the Air Force Safety Center deployed the automated system to reduce the cycle time in resolving safety issues by transforming time-consuming manual classification to the automated system (Wolfe et al., 2008).

IBM has developed a semi-automated system for structured query language (SQL). InfoSphere Information Server actually applies tags when user provides SQL statements.

According to IBM, "SQL Meta Tags are platform-independent SQL functions that are supported by the Dynamic Relational Database stages. At runtime, these tags are translated into native database-specific SQL functions of the backend database" (IBM, 2012). Thereafter, these tags remain in the database.

Many email systems automatically append messages with metadata. Microsoft Outlook can store emails with Message (MSG) or Personal Storage Table (PST) file name extension. Both formats preserve metadata (Losey, n.d.). Outlook commands 43 percent of the email client market (Litmus, 2012).

Dublin Core Metadata Initiative (DCMI) establishes semantic metadata standards. Through communities, task groups, forums, conferences and a website, DCMI sets detailed standards that support interoperability. Communities include Knowledge Management, Science & Metadata, and Localizations and Internationalization. Task Groups include Metadata Provenance, Vocabulary Management, and Abstract Model Review (Dublin Core Metadata Initiative, n.d.). The Abstract Model might be a means to extend semantic taxonomies into the realm of valuation of information. It is flexible enough to allow encoding in HTML meta tags, Resource Description Format (RDF), and XML (Dublin Core Metadata Initiative, n.d.).

In 2005, Jin has proposed the use of MetaXQuery, which would automate the conversion of low-level algebraic expressions in a database management system (DBMS) into XML (Jin, 2005). XML is likely to play a lead role in valuation of information. WC3 defined XML, but did not invent the concept. Standard Generalized Markup Language (ISO 8879) – which is somewhat simple and flexible – was already being used. XML was built on the idea of a markup language to support the growing electronic publishing industry, much as IPTC had for news data. The use of XML is growing as the web grows" (WC3, 2012).

As an example of XML, <animal>**dog**</animal> illustrates that the tag "animal" defines how the term "dog" should be classified. These descriptors can add value to the data, especially as data migrates and is combined with other data. XML need not have any proprietary storage format. Therefore, XML can operate with databases management systems (DBMS). Supported application programming interfaces (APIs) include BaseX, eXist, MarkLogic Server, and Sedna (BlurtIt, 2012).

There are four common models in distributed database systems: Relational Model, Functional Data Model, Entity Relationship Model, and Semantic Database Model (Ram & Liu, 2006).

The Relational Model is based on the idea that information is stored in separate tables that relate to one another. They are linked and data is referenced between tables (BlurtIt, 2012). Relational databases capable of operating with ISO XML include Microsoft SQL Server and PostgreSQL, as well as IBM DB2 and Oracle Database (BlurtIt, 2012). Shanmugasundarum et al. (1999) explored the feasibility of querying over documents tagged with XML. Their research confirms that two methods work in the relational model. 1) XML Schemas can be applied to Excel files and imported into DBMS; 2) Document Type Descriptors (DTDs) enable XML documents to be parsed and loaded into tuples. Semi-structured data then needs to be translated to Lorel or XML query language on top of SQL queries over the relevant relational data. They used IBM DB2 to validate this method (Shanmugasundarum et al., 1999).

Functional Data Model searches have two parts: the query and the constructor. The query is the expression to be evaluated above the information. The constructor part wraps query results and forms the XML output, based on variables, constants, tuples, projections,

applications and abstractions using Functional Data Manipulation Language (FDML). Lambda Calculus is used to combine functions expressed in database entities (Coronet, 2012). Loupal ran tests on the Functional Data Model and confirmed DTD constructs can be applied without breaking the validity of XML schema. This approach is suitable for queries and updates of XML documents (Loupal, 2012).

Entity Relationship Model is based on entities (e.g., person) and relationships (i.e., logical connections) (Coronet, 2012). Fundulaki and Marx (2012) have tested this model using an ontology-based mediator, which enables querying of heterogeneous XML resources. Local-As-View and Global-As-View (GAV) are the two most common approaches to mapping schemas at the global and source levels. GAV is the basis for the model validated by the authors, which they describe as "query rewriting using views." These can be executed using RDF Schema and object-oriented architecture. As a result they are able to integrate web data into data warehouses with an expressive mapping language, using the ST_YX prototype. The authors note that the mediator is a rewriting algorithm that "examines the query variables and finds the mapping rules which provide the answers for them" (Fundulaki, 2012).

The Semantic Database Model (SDM) is a structure designed to identify the meaning of an application. SDM Schema offers an accurate documentation and communication medium to identify what is really used in the database, and is commonly used to produce an interface for non-programmers (Hammer, 1981). Ram and Liu's (2006) W7 approach is based on the Semantic Database Model, which includes the use of an algebraic analyzer, polygen operation interpreter and query optimizer.

SQL Server has the capacity to query XML metadata using SQL (StackOverflow, 2012). Graziano notes that the SQL-92 standard included information schema views. Functions that support this capability include ObjectProperty and DataProperty (Graziano, 2003). With SQL Server 2005, Microsoft introduced metadata views such as sys.tables and sys.procedures as well as INFORMATION_SCHEMA views. Microsoft secures metadata views on a per row basis, and will not allow users to see the metadata, unless they have authorized access (Beauchemin, 2012). If they do have access and conduct a query, they may find metadata errors. This provides an opportunity to report incorrect metadata that were automatically tagged earlier in the process.

Building Block Three: Learning Machines

Correcting metadata errors at scale is an important part of this model. Tuzhilin, Liu and Hu have created a model management system that tests the quality of the data and meta tagging, using the query language ModQL, which is an object-relational dialect of SQL99. (Tuzhilin et al., 2012). It requires the expertise of a Data Scientist to identify flaws and recommend changes as needed. Ideally, a different person would make the changes, perhaps a Metadata Curator.

Machine learning can be applied to the relatively small dataset generated by this administrator. Predictive coding is essential to this concept. This can be used to improve the way metadata is automatically tagged. Machine learning – a form of artificial intelligence – can make predictions based on the new data through training labels. As a result, classifiers train the system on correct parameters. The Curator has the option of using four methods to do this: classification or categorization, regression, clustering or dimensionality reduction. Classification is a form of supervised learning that is discrete and is most appropriate for teaching automated systems to apply meta tags (Nilsson, 1998)

The key to success is to not only correct new incoming data, but to also retroactively update data across the enterprise. This requires ongoing data mining to identify and extract relationships that may not have been tagged correctly during an earlier stage. Therefore, machine learning is capable of helping an enterprise to evolve (Nilsson, 1998). This can be very valuable for organizations that feel locked-into legacy systems.

The Curator can work within a set of policies and standards that also determine what can be aggregated and shared with interested parties – such as investors – without compromising confidentiality or privacy.

TRACKING AND REPORTING

Many Certified Public Accountants (CPAs) who work with public companies are already familiar with XML in the form of eXtensible Business Reporting Language (XBRL). The industry consortium XBRL US sets the standards – based on XML. Each financial statement is meta tagged with a vocabulary that included descriptions, units and currency. The classifications are designed to be easy for users, such as analysts and investors, to understand and work with. Because of its extensibility, it can be customized and expanded to suit the needs of different industries. "Extensions" provide flexibility (XBRL, 2012).

EdgarDashboard.XBRLcloud.com allows investors to see the adoption rates of public companies. To date, none have reached 100 percent compliance. It reveals errors, warnings, inconsistencies, failure to comply with best practices and other non-standard information (XBRL Cloud, 2012).Extensible data is already a work in progress, and companies are grappling with the best ways to use the technology.

For the last decade, PricewaterhouseCoopers (PWC) has been working with companies to incorporate sustainability initiatives into XBRL. For example, ISO 14000 series reports on processes for producing a product negatively affects air, water or land. SA 8000 audits and reports on employee rights and working conditions. In other words, XBRL has the capacity to measure and report "off balance sheet" data, such as intangible (yet quantifiable) capital (PWC, 2012). Companies capable of reporting the value of their intangible information assets may have an advantage with investors – to reduce Market Value at Risk (MVR).

CONCLUSION

A set of global taxonomic standards will need to be created for use in meta tagging. Global standards can be used by each company to customize their own standards. The two most logical organizations to collaborate on this initiative would be XBRL US and Dublin Core. To scale-up the concept, automated tagging is required to save the time of staff, who are likely to feel they have better things to do than tag every single piece of datum. In addition, to define what is needed to encourage compliance, it would be useful to obtain the buy-in, support and endorsement of the leading accounting firms, the Financial Accounting Standards Board (FASB), Securities and Exchange Commission (SEC) and other influential organizations.

Multiple dimensions for each piece of datum makes the value of information complex and ever changing, but systems already exist that could be deployed to query new metadata. If organizations treat each piece of datum as an asset, liability or risk, they may be able to develop the processes to measure its value that translate to MVR.

With the proper dashboard and reporting mechanisms in place, senior management

will have the ability to see the total aggregate of information value. They will also be able to drill-down to highly detailed views. This new information is essential in calculating MVR. It is also essential in prioritizing what matters most and least.

Based on metrics that CFOs can accept, there may be new opportunities to secure intangible capital of interest to hackers, rogue states or competitors. The most valuable information assets can remain inside the organization's most secure systems. This information has service potential or future economic benefits. The least valuable data can be retained in less expensive cloud solutions. This model gives organizations a powerful new way to quantify what investors and hackers already know: Information is incredibly valuable.

REFERENCES

Beauchemin, B. (2012). "SQL server 2012 security best practices – Operational and administrative tasks." Microsoft SQL Server 2012.

BlurtIt. (2012). "What is relational model In DBMS?" Retrieved from http://sql-databases.blurtit.com/q5030694.html.

Coronet. (2012). "Functional data model." Retrieved from http://coronet.iicm.edu/dm/scripts/lesson06.pdf.

Csaplar, D. (2012). "Cloud Storage Gateways – Large Enterprises are Learning what SMBs Already Know." Retrieved from http://blogs.aberdeen.com/it-infrastructure/cloud-storage-gateways-large-enterprises-are-learning-what-smbs-already-know/.

Daily Mail Reporter. (2012). "The camera that 'prints out' what it sees – As well as takes a photo." Retrieved from http://www.dailymail.co.uk/sciencetech/article-2135341/The-camera-prints-sees--taking-photo.html.

Dublin Core Metadata Initiative. (n.d.). "DCMI work structure." Retrieved from http://dublincore.org/groups/#communities.

Federal Geographic Data Committee. (2012). "Content standard for digital geospatial metadata." Retrieved from http://www.fgdc.gov/metadata/csdgm/.

Fundulaki, I. & Marx, M. (2012). "Mediation of XML data through entity relationship models." Bell Labs, Lucent Technologies and INRIA-Rocquencourt and Institute for Logic Language and Computation, University of Amsterdam.

Gartner (2012). "Gartner says master data management is critical to achieving effective information governance." Retrieved from http://www.gartner.com/it/page.jsp?id=1898914

Gartner. (2012). "Information governance." Retrieved from http://www.gartner.com/it-glossary/information-governance/

Google Webmaster Central Blog. (2009). "Google does not use the keywords meta tag in web ranking." Retrieved from http://googlewebmastercentral.blogspot.com/2009/09/google-does-not-use-keywords-meta-tag.html.

Graziano, B. (2003). "Using metadata." SQL Team. Retrieved from http://www.sqlteam.com/article/using-metadata.

Hammer, M. & McLeod, D. (1981). "Database description with SDM: A semantic database model." ACM Transactions on Database Systems. Retrieved from http://ece.ut.ac.ir/dbrg/seminars/AdvancedDB/2007/Rostami%20hosein%20-%20Zohdi%20alireza/Report2/Articles/Database%20Description%20with%20SDM%20-%20A%20Semantic%20Database%20Model.pdf.

IBM. (2012). "SQL meta tags." Retrieved from http://publib.boulder.ibm.com/infocenter/iisinfsv/v8r7/index.jsp?topic=%2Fcom.ibm.swg.im.iis.conn.drs.doc%2Ftopics%2FDRS049.html.

James, P. (2010). "How Much Data Do Americans Consume Each Day?" Good Design. Retrieved from http://www.good.is/post/how-much-data-do-americans-consume-each-day/.

Jin, H. (2005). "A framework for capturing, querying and restructuring metadata in XML data." Dissertation, Washington State University.

Lastowka, G. (1999). "Search engines, HTML, and trademarks: What the Meta for?" Ritgers School of Law. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=913990.

Litmus. (2012). "Email client market share." Retrieved from http://litmus.com/resources/email-client-stats.

Logix4U. (2012). "Introduction to Storage Systems." Retrieved from http://www.logix4u.net/component/content/article/23-introduction-to-storage-systems/29-storageintro1.

Losey, R. "Email metadata and production." (n.d.). Retrieved from http://floridalawfirm.com/msg.html.

Loupal, P. (2012). "Updating typed XML documents using a functional data model." Dept. of Computer Science and Engineering, Czech Technical University. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.85.8496.

MacVittie, L. (2012). "Big data: Why it's really an architecture challenge." ZDNet. Retrieved from http://www.zdnet.com/big-data-why-its-really-an-architecture-challenge-7000006699.

Mattsson, U. (2012). "How to prevent internal and external attacks on data – Securing the enterprise data flow against advanced attacks." SSRN. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1144290.

Media/Outreach. (2012). "A brief history of meta tags." Retrieved from http://mediaoutreach.com/2008/11/a-brief-history-of-meta-tags/.

New York State Bar Association – Committee on Professional Ethics Opinion 749. (2001). Retrieved from http://www.nysba.org/AM/Template.cfm?Section=Ethics_Opinions&TEMPLATE=/CM/ContentDisplay.cfm&CONTENTID=6533.

Nilsson, J. (1998). "Introduction to machine learning." Retrieved from http://robotics.stanford.edu/~nilsson/MLBOOK.pdf.

OCLC. (2012). "How one library pioneer profoundly influenced modern librarianship." Retrieved from http://www.oclc.org/dewey/resources/biography/.

Online Etymology Dictionary. (2012). "Meta." Retrieved from http://www.etymonline.com/index.php?allowed_in_frame=0&search=meta&searchmode=none.

Online Etymology Dictionary. (2012). "Tag." Retrieved from http://www.etymonline.com/index.php?allowed_in_frame=0&search=meta&searchmode=none.

Perlman, A. (2009). "The legal ethics of metadata mining." Akron Law Review, Soffolk University, SSRN. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id_=1472712

Photo Meta Data. (2012). "Metadata history: Timeline." Retrieved from

http://www.photometadata.org/META-Resources-Metadata-History-Timeline.
PWC. (2012). "Closing the loop on sustainability information." Retrieved from http://www.pwc.com/us/en/technology-forecast/2011/issue4/features/feature-technology-enabling-sustainability.jhtml.
Ram, S. & Liu, J. (2006). "Understanding the semantics of data provenance to support active conceptual modeling." Proceedings of the Active Conceptual Modeling of Learning Workshop (ACM-L 2006) in conjunction with the 25th International Conference on Conceptual Modeling (ER 2006).
Shanmugasundarum, J., Tufte, K., He, G., Zhang, C., DeWitt, D., & Naughton, J. (1999). "Relational databases for querying XML documents: Limitations and opportunities." Department of Computer Sciences, University of Wisconsin-Madison, Proceedings of the 25th VLDB Conference. Retrieved from http://www.cs.cornell.edu/people/jai/papers/rdbmsforxml.pdf.
StackOverflow. (2012). "How to do SQL query for XML data (in SQL Server)?" Retrieved from http://stackoverflow.com/questions.7483745/how-to-do-sql-query-for-xml-data-in-sqlserver.
Tuzhilin, A., Liu, B., & Hu, J. (2012). "Building and querying large modelbases," Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1281307##.
United States Geological Survey. (2012). "Frequently asked questions on FGDC metadata." Retrieved from http://geology.usgs.gov/tools/metadata/tools/doc/faq.html#q1.1.
University of Illinois. (2012). "Dewey decimal system - A guide to call numbers." Retrieved from http://www.library.illinois.edu/ugl/about/dewey.html.
WC3. (2012). "Extensible Markup Language (XML)." Retrieved from http://www.w3.org/XML/.
Wolfe, S., Sanchez, D., Chaple, S. (n.d.). "Automated metatagging, taxonomy management and auto-classification in an enterprise environment." 20th International Conference on Systems Research, Informatics and Cybernetics. Retrieved from http://www.conceptsearching.com/web/userfiles/file/InterSymp%202008%20AFMS.pdf
Wolfe, S., Sanchez, D., & Chaple, A. (2008). "Automated metadata tagging, taxonomy management and auto-classification of information in an enterprise environment." Preconference Proceedings: 20th International Conference on Systems Research, Informatics and Cynernetics, InterSymp 2008, Focus Symposium on Intelligent Software Tools and Services, 2008.
XBRL Cloud. (2012). "EDGAR dashboard." Retrieved from https://edgardashboard.xbrlcloud.com/edgar-dashboard/dashboard.do.
XBRL US. (2012). "Fact sheets." Retrieved from http://xbrl.us/Learn/Pages/FactSheet.aspx.

APPENDIX A

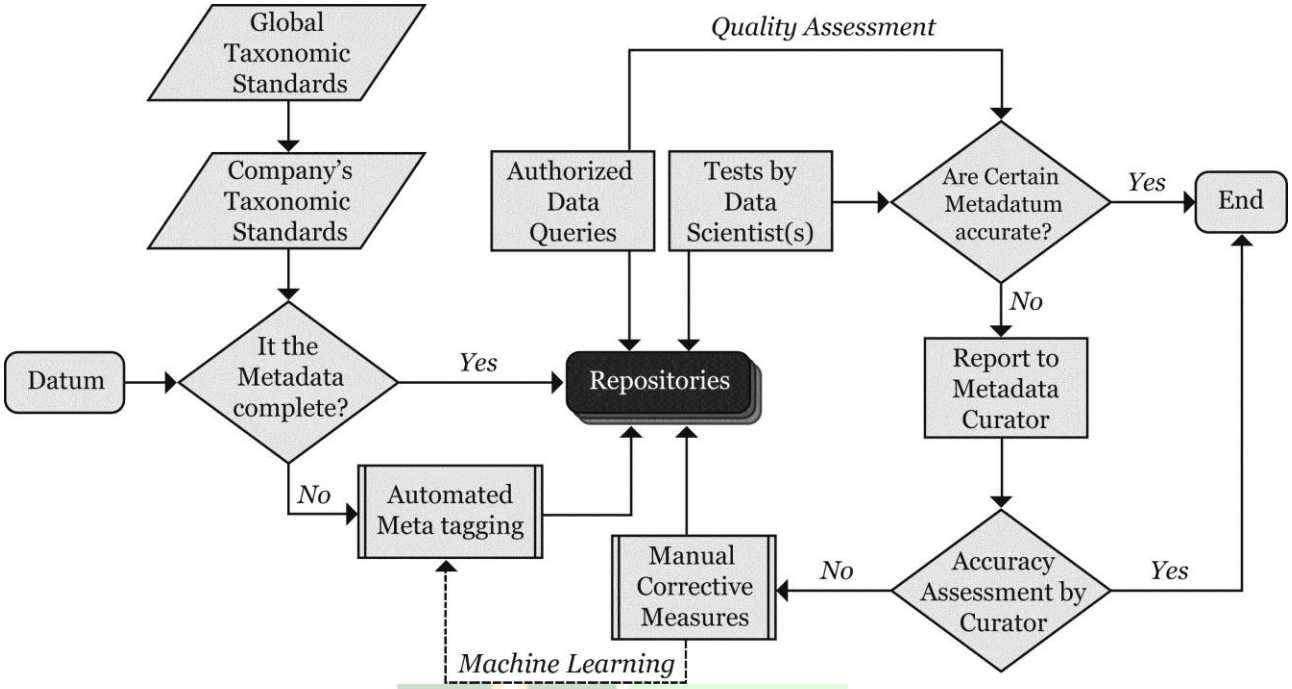

Figure 1. This information workflow diagram illustrates the proposed model that will enable large volumes of data to be automatically meta tagged with predictive coding and revised using machine learning. An administrator is required to supervise changes and machine learning. They could be considered a "Curator."