TRUSTED DATA MARKETS AND THE ROLE OF DATA QUALITY

(Practice-oriented Paper)

David Becker
The MITRE Corporation
dbecker@mitre.org

Abstract: Optimization of information exchange among data producers and consumers in external settings between business partners, and internally within large organizations, is becoming increasingly important. Historically, information was exchanged informally based on established relationships through direct interaction. However, as more information exchanges move online involving digital products, challenges have arisen: parties to an exchange don't know each other, data needs are time sensitive, multidimensional and dynamic, quality of data assets is unknown or difficult to ascertain, and technology differentials are significant. Many techniques have arisen to moderate these problems, but significant inefficiencies continue making exchange costly and time-consuming. Electronic marketplaces, which have proven successful at minimizing transaction costs of physical product exchanges, are now being applied to information products. But current electronic data markets have important shortcomings that have not been adequately addressed. Trusted Data Markets provide a framework for addressing those shortcomings, in particular trust issues associated with data quality.

Key Words: Data Markets, Trust, Information Quality

PROBLEM

As more & more information exchanges involving digital products move online, challenges have arisen. For example parties to an exchange frequently don't know each other. Data needs are often very complex, being time sensitive, multidimensional and/or dynamic. The semantics of the expression of these data needs by potential users differ from the descriptions and terms of use of the available data sources. The quality of data assets is almost always unknown or very difficult to ascertain. There are big differences between the technological capabilities of the buyers and sellers. As a result, set-up and transaction costs are significant for both data consumers and producers

Various efforts to moderate these challenges have actually added significantly to the costs of setting up and conducting transactions: e.g., elaborate authentication & authorization schemes, complex acquisition processes, agile programming techniques for rapid development of access services, Web 2.0 ontologies & vocabularies, etc. Thus, large inefficiencies persist, continuing to make information exchange a costly and time-consuming activity. In some cases the challenges remain unaddressed or inadequately mitigated.

ONE POSSIBLE SOLUTION – MARKETS

One approach to addressing many of the challenges of data exchange is to adopt a market-based solution. In a free market, if you let buyers and sellers trade with each other (within an auction context), the bids and asks will quickly (within the auction timeframe) converge on a single price, which is the price where supply and demand meet. [Surowiecki p104]. Markets tend to move towards prices which balance the quantity supplied and the quantity demanded, such that the market will eventually be cleared of all

surpluses and shortages (excess supply and demand) [Wikipedia –Market clearing"]. Such a market will maximize the group's total gain from the trading despite conflicting self-interests, and imperfect, incomplete information. A well-functioning market will make everyone better off than they were when trading began. Harnessing the competitive pressure of commercial markets is increasingly being used as a policy tool to reduce the costs of securing outcomes and to create greater flexibility for delivering outcomes [Designer Carrots].

Data Markets

Markets are historically associated with the buying and selling of tangible products. However, markets have also arisen to address the buying and selling of intangible items, such as stocks. For our purposes, markets are also emerging to support the buying and selling of information products, what we call data markets. Bill Day (quoting Grant Nestor of Factual) has defined a data market as –a destination where data is exchanged for other data, money, or things of value" [Day, Part 1].

One way to characterize electronic data markets is to look at their focus and how they present their available datasets for consumption by developers, analysts, and other users [Day, Part 1]:

- Data catalogs: Markets that pull together links to various datasets; the data may be hosted in the market's own storage or it may be linked to elsewhere. Catalog markets are meant to make it easier to locate datasets of interest, but the data itself may not always be as "fresh" as the data provided by real time feeds.
- Real-time feeds: Services that provide direct access to constantly updating streams of relevant data e.g., Twitter feeds. This type of market offers super-fresh data, but can also provide an overwhelming volume of data to store and process if not careful.
- Free public data sources: Often mandated by governments and NGOs, these provide access to useful data, but the data may be poorly structured or "dirty," making it more difficult for a developer to use.
- Graphics-oriented services: Services meant more for analysts than developers—heavy on built-in visualization tools and spreadsheet support, but often lacking in programmatic (API) access.
- Internal Data Markets: Internal Markets consist of groups within an organization exchanging goods, services and information. Internal data markets are important for achieving internal efficiencies and —cause" objectives [Koronios]

One traditional area for electronic data markets involves business information [Simba]. This includes brokerage information, financial news and information, credit information, legal, tax and public record information, health care information, general news and research information, marketing information, and other online information services. There are many commercial business information vendors who sell their data sets, reports and analyses to corporations and individuals for a profit: Reuters, Thomson, Factset, Dow Jones, MarketWatch, Acxiom, Bloomberg, ChoicePoint, Dun & Bradstreet, Lexis-Nexis, IRI, Springer, Equifax, Emdeon (WebMD), ValueLine , etc.

Another area for data markets revolves around personal data. This can be broken down into 3 parts. First are the primary data markets. In many cases, appropriately processed personal data are valuable market assets since they can be used to improve services, increase sales, etc. [Kiayias]. Next are secondary data markets: It is often important to transfer original personal data from the primary market where they are collected to a secondary market where other parties will further process them. Secondary processing of the original (much more private) data frequently involves making the data anonymous. The third part consists of black markets for stolen IDs and credit cards: Criminals who steal personal data often don't use it themselves. Instead, they put it up for sale on one of the many online markets [Sutherland].

A third major area for data markets involves government data dump sites. A huge amount of information is available from federal, state and local governments. While these sources are typically free and frequently of dubious quality, they provide a rich and varied assortment of data sources: e.g.,

census/demographic, environmental, patent, financial filings, medical, etc.

Finally there are a host of other areas where data markets have been or are being set up: music & video, smart phone apps, social networking information, satellite data, residual government insurance markets, various state (e.g., Maryland) and regional energy markets, air traffic flow management, bumping on oversold flights, etc. Essentially wherever data goes digital, data markets are emerging.

Electronic Data Marketplaces

Many web sites have become official destinations for conducting online retail transactions. These electronic marketplaces (e.g. eBay) have proven successful at reducing transaction costs in the exchange of physical products. Electronic marketplaces are also being successfully applied to the exchange of information products: These are called electronic data markets. An electronic data market takes the data market concept online by implementing the exchange of data assets through the internet.

Some examples of the new crop of electronic data market vendors [Gislason, Day] include the following companies: DataMarket, Trimetric, Google Public Data, Wolfram|Alpha, Infochimps, Factual, Kasabi, Freebase (purchased by Google), Windows Azure Marketplace DataMarket (formerly code named "Dallas"), BuzzData, and Socrata's OpenData. It is the trust and transactional efficiency aspects of these electronic data marketplaces and the data markets described above that are the subject of this paper.

BACKGROUND

In order to evaluate the specifics of the challenges as they relate to markets, we need to look a little closer at how they are set up and operate.

Markets

The definition of a market from Wikipedia [Wikipedia "Market"] is as follows: Any structure that allows buyers and sellers to exchange any type of goods or services. A market exists in order to enable the exchange of rights (cf. ownership) to products (goods or services).The exchange of goods and services is a transaction. There are two primary roles in markets: buyers and sellers. Markets allow any tradable item to be evaluated and priced. Bids represent the highest price a buyer/customer is willing to pay for a good, while asks (offers) represent the price a seller/supplier is willing to accept for that particular good. As the valuation of the consideration given in exchange for transfer of ownership, price forms the essential basis of commercial transactions. Price may be fixed by a contract, left to be determined by an agreed upon formula at a future date, or discovered or negotiated during the course of dealings between the parties involved. In commerce, price is determined by what (1) a buyer is willing to pay, (2) a seller is willing to accept, and (3) the competition is allowing to be charged [BusinessDictionary, "Price"]. The marketplace facilitates trade and enables the distribution and allocation of resources in a society. A market emerges more or less spontaneously, or is constructed deliberately by human interaction.

Market Structure

Market Structure [Koronios] consists of those components of various types that are arranged in specific combinations to make up a market that achieves the objectives of exchange between customers and suppliers (see Figure 1). For a data market, the products are the data and information assets produced by one party and consumed by another party. The customers are the various groups and individuals (with a willingness & ability to buy) that are looking for data assets which meet their consumption needs. Suppliers are the various providers (with a willingness & ability to sell) who own or fund the creation and maintenance of the data assets. The market governing body is the company or organization responsible for setting up and operating the market.

An auction is the basic process by which customers are matched with suppliers, price is discovered or negotiated, and business transactions are set up, processed, and completed. Auctions can take many forms (fixed price, forward, reverse, Dutch, sealed bid, etc.). The marketplace is the infrastructure through which customers are matched with suppliers during an auction. Various market mechanisms are provided as tools to support to the exchange process, such as product descriptions, quality, price, and terms under which data assets may be bought and sold (traded with value). Market regulation comprises all of the laws, rules, and customs that govern the activities in the marketplace & its mechanisms, and the constructs which enforce them.

Frequently there are various intermediaries or middlemen who participate in the market to help address market inefficiencies. They are essentially the grease for the wheels of the market. These include wholesalers, retailers, value-added resellers, agents, brokers, and contractors. In addition the market typically has a governing body that is responsible for setting up and operating the market.

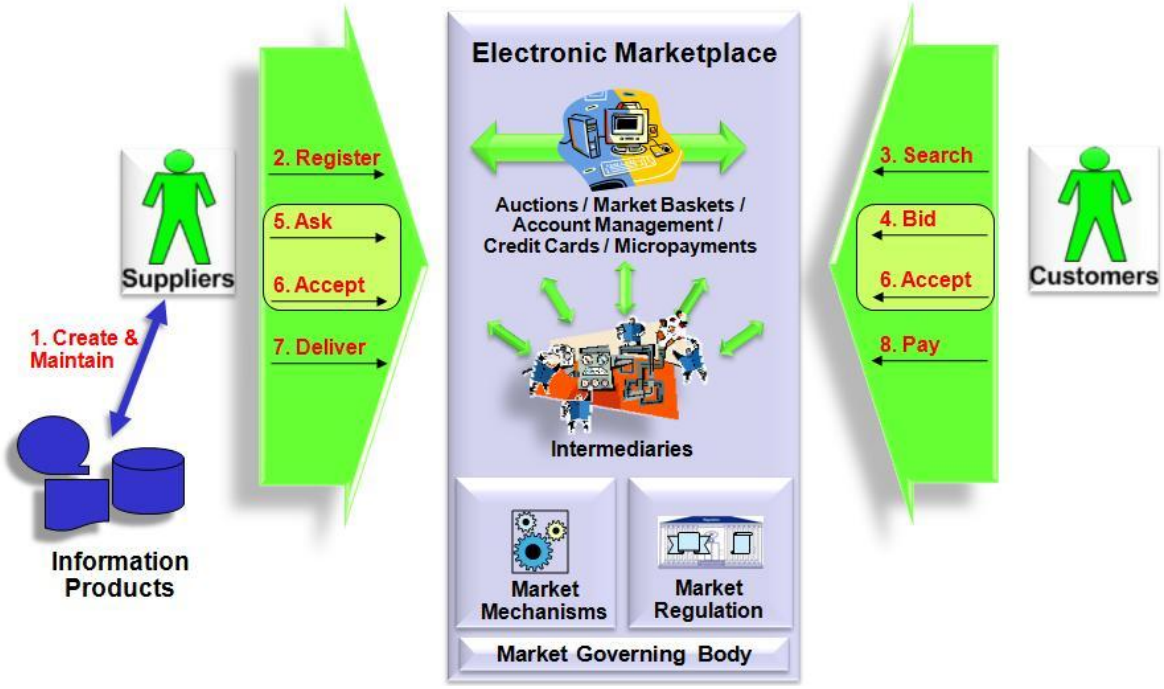

Figure 1 - Electronic Data Markets

The process associated with the electronic data market typically proceeds as follows. It starts with a data asset or information product that is created and maintained by a supplier. If the supplier is interested in making this product available for acquisition by various customers, they can register the product within the electronic marketplace. This would include all information which properly describes the product and various other pertinent characteristics. A consumer interested in products of a given type would conduct a search on the products registered through the electronic data market. The negotiation involved in the next step can proceed in many different ways, but would involve a bid from the consumer side and an ask from the producer side. If both parties reach agreement on the product to be provided by the producer and the consideration to be provided by the consumer, then a formal acceptance can be made on each part. Delivery or provision of access to the product can then be arranged, and the consumer can receive or retrieve the product. Upon satisfactory receipt of the product, the consumer would then provide his agreed

upon consideration or payment to the producer, who would provide acknowledgement.

Shortcomings of Electronic Data Markets

One of the principal shortcomings of current electronic data markets involves deficiencies related to trust. The need for buyers to be able to trust sellers has been heightened by the use of the internet, which introduces a large amount of anonymity and physical separation. Differences in prices among Internet retailers can often be explained by differences in trust. For example, buyers will usually pay much less for products from unknown sellers of products whose quality they cannot verify. Buyers willingly pay more to reduce their uncertainty and improve trust. Some retailers are trusted more by consumers and can, therefore, charge higher prices. Well-designed markets have a variety of mechanisms, formal and informal, to overcome this problem by ensuring there are incentives for market participants to be honest [McMillan p54]. Marketplace confidence rests on rules and customs that provide such incentives.

Another shortcoming of electronic data markets involves market friction that can arise. Market frictions constitute anything that prevents markets from developing and working properly, and that imposes costs or restraints on business transactions [Frictionless Markets; Designer Carrots]. In data markets, market frictions arise when the different components set up for the market don't operate smoothly or require excessive time, effort and resource to function. Market frictions still exist in many electronic data markets, frequently related to trust and data quality that in turn drive higher costs. These will be discussed in detail below.

Electronic Data Markets & the Hierarchical Framework of Data Quality

As we will discover, many of the problems of information exchange and electronic data markets can be traced back to data quality problems. A common definition of high-quality data is "data that are fit for use" [Juran, Section 34.8, Wang]. This means that different aspects of use must be considered. A good way to characterize the different types of data quality that are applicable to different components of the electronic data market and that address the different aspects of use is to apply the hierarchical framework (See Figure 2) developed by Wang and Strong [Wang].

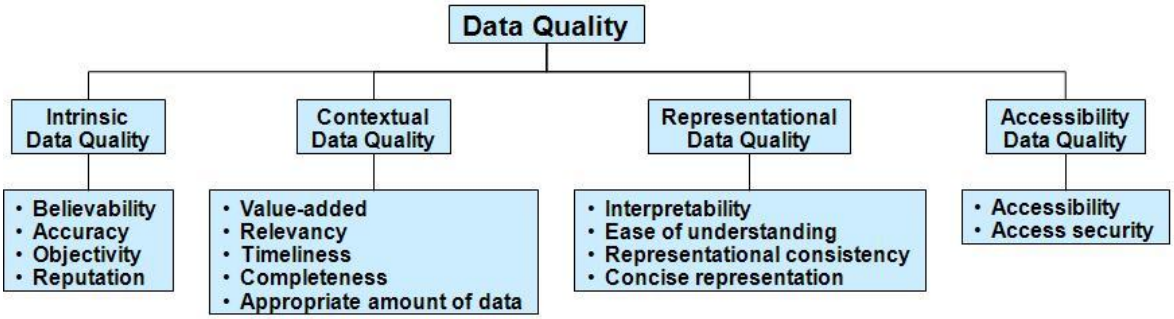

Figure 2 - Hierarchical Framework of Data Quality

In this framework, the major categories of data quality are: intrinsic, contextual, representational, and accessibility. Each of these categories can contain specific DQ dimensions. Intrinsic DQ consists of the accuracy, objectivity, believability, reputation dimensions. Contextual DQ consists of the relevancy, value-added, timeliness, completeness, and amount of data dimensions. Representational DQ consists of the interpretability, ease of understanding, concise representation and consistent representation dimensions. Accessibility DQ consists of the accessibility and access security dimensions.

All of these categories are applicable to electronic data markets in one form or another. Data quality problems in each of these dimensions can contribute to problems in establishing a foundation for trust among the different market participants, and in reducing various market frictions.

Hypothesis

So, the basic hypothesis of this paper is that if we can set up a framework for identifying and organizing the shortcomings of electronic data markets, then this framework can be used to analyze and address these shortcomings in a systematic way as they occur in practice for electronic data markets that are already operating or have operated in the past, or in the design and implementation of new or proposed electronic data markets.

To date our investigations have found no such framework. While this paper attempts to propose a framework, it does not attempt to evaluate its adequacy. This will be the subject of future work.

Furthermore, because many shortcomings are the direct result of problems with data quality, if a direct mapping can be made between the framework for identifying and organizing the shortcomings of electronic data markets and the hierarchical framework of data quality introduced above, then the huge body of data quality management knowledge built up around the DQ framework can be brought to bear in an organized fashion for the analysis and improvement of electronic data markets.

SOLUTION APPROACH – TRUSTED DATA MARKETS

One way to address the current shortcomings of electronic data markets is to set up a ―trusted data market". A trusted data market (See Figure 3) is an electronic data market framework for online data suppliers and customers that applies proven market techniques for finding, evaluating, acquiring access to, and exchanging information assets while providing a structure for addressing the need for trust, and significantly reducing market frictions particularly related to data quality.

Figure 3 - Trusted Data Market Framework

THE NEED FOR TRUST

There have been a number of studies that have established the need for trust to achieve overall success in various collective exchange situations [Surowiecki p140], such as the ultimatum game, public-goods games, and analyses of getting people to pay taxes. Persistent concerns with trust can become a serious disadvantage for a market [Surowiecki-2 p166].

An example is eBay where people may have something extra to worry about. While all evidence suggests that fraud is unusual on eBay, there have been a few high profile incidents. eBay does little to vet sellers

in line with its desire to create and open, relatively unregulated bazaar. In contrast, people don't worry about Amazon.com. Amazon sets high hurdles for companies to be able to sell in its Marketplace in the first place, and then stringently monitors companies once they are allowed in.

These types of findings are directly applicable to all electronic data markets. As stated previously, the need for buyers to be able to trust sellers has been heightened by the use of the internet.

There are three kinds of trust that must be addressed:

- Trust in the governing body – that it will develop and exercise the rules fairly and apply them consistently in everyone's best interests
- Trust that justice will be served – that the governing body will find and punish the guilty and avoid punishing the innocent
- Trust between buyers and sellers – that they will generally do the right thing and live up to any reasonable obligations

Trust in the Governing Body – Governance Trust

Market participants must feel that the governing body is operating in everyone's best interests [McMillan p54]. They must believe that the governing body has developed a fair set of rules not slanted in any parties favor. They must also believe that the rules are applied in a consistent manner, and that the operation of the market does not consume an excessive amount of resource (performance and overhead).

The degree of trust in a political or economic authority is often critical in determining the success of a market [Wikipedia —Market System"]. As a result more parties will participate in the market, they will participate more frequently, they will be more satisfied with the outcomes, and there will typically be fewer disputes.

There are a number of governance trust mechanisms.

- Respect – That the governing authority is credentialed, competent, skilled and experienced in the governance of the particular market or similar markets
- Market rules – That there is clear definition and publication of the market rules, processes and procedures to be followed by different market participants
- Operations monitoring – That sufficient oversight will be provided to ensure various activities in the market proceed smoothly and efficiently
- Adjudication of disputes – That adequate constructs exist to provide parties a fair and open hearing of disagreements, and that binding decisions can be quickly rendered
- Overhead – That the cost of participating in the market is reasonable with respect to the benefits to be derived

Trust That Justice Will Be Served – Regulation Trust

This kind of trust involves the fair and consistent policing and enforcement of market rules and social conventions. Regulation trust should address the believability, objectivity and reputation dimensions of the intrinsic DQ category.

There are a number of regulation trust mechanisms. One of these is reputation. A provider's reputation (such as a brand name or an authoritative source designation) can convey needed levels of assurance. Systems that provide a means of establishing and maintaining a reputation are

For example, how does eBay protect its buyers from fraud or misrepresentation by sellers? Buyers cannot check the quality of the merchandise. As a solution, eBay provides a reputation system as a mechanism to prevent fraud. It allows regular sellers to establish a reputation for reliability and quality. [McMillan p78] After an auction eBay asks the buyers to rate the seller, and then posts the ratings and comments online for anyone to see. A seller is given a score that counts the number of favorable and negative comments

received. This simple device works. eBay claims that fraud affects less than 1% of its auctions. A reputation for honest dealing is very valuable. Bids will typically go significantly higher when the seller has a high rating. Furthermore, this mechanism is self-policing. Once set up, it does not require significant oversight from eBay.

Another regulation trust mechanism is the traditional investigator. Traditional investigators were used by eBay to supplement its reputation mechanism with more formal methods of fraud prevention. It hired traditional investigators to track down thieves and con artists using the site.

Standard audits can be conducted periodically by the governing body or a regulatory agency to monitor the level of compliance. Sometimes these are dictated by legislation or regulations such as Sarbanes-Oxley (SOX) or government CFO compliance. Compliance requirements can frequently drive design and implementation activities, and can result in the capture and handling of significant amounts of metadata. Organizations often must demonstrate the enforcement of separation of duties and avoidance of conflicts of interest.

Another regulation trust mechanism involves monitoring and data mining. In this case, manual or automated inspection and analysis of transactional activity is performed in order to detect unusual or suspicious behavior and trends.

Trust Between Buyers and Seller – Relationship Trust

Market participants must believe that buyers and sellers will generally do the right thing and live up to any reasonable obligations. However, parties to an exchange frequently don't know each other. There is very little basis for trust. Any problems in the trust relationship between buyers and sellers can increase risk and significantly affect prices or value. So a number of mechanisms have been put in place to address this issue.

Because prior relationships can often provide the needed levels of assurance, mechanisms for buyers and sellers to connect with one another to begin to establish relationships are very valuable. For example: organizational web sites, chat rooms, blogs, fact sheets, contact lists, reference accounts, FAQs, Wikis, trade and industry groups, and other mechanisms.

Other approaches for dealing with relationship trust issues involve external companies or organizations. Credit bureaus collect information from various sources (creditors, lenders, utilities, debt collection agencies, and public records) and provide it to sellers, for a fee, to assess credit worthiness of buyers [Wikipedia —Credit bureau"]. Also, escrow companies will hold a buyers payment, for a fee, until the buyer has received the item satisfactorily.

There are several security aspects of Relationship Trust [Wikipedia —Information security"]:

- Authentication: Buyers and sellers are who they say they are, and information assets and transactions are genuine
- Authorization: Buyers and sellers can legally engage in market activities at levels necessary to successfully conduct business (also called access control)
- Confidentiality: Information assets and market information is made available or disclosed only to those who are authorized
- Integrity: Messages and electronic transactions have not been actively modified in any undetectable way
- Non-repudiation: Buyers and sellers cannot renege on agreements & parties to a transactions cannot deny their particular sending or receiving roles

There are a host of tools & security trust mechanisms that have been developed over the years to address these different aspects of security. The extent and intensity of use of security trust mechanisms is dictated

by the sensitivity of the data assets, the need for anonymity, and risk tolerance of the involved parties.

Quality Dimensions Related to Trust

With respect to the hierarchical framework of data quality, several of the categories and dimensions come into play when overcoming trust issues. The different security mechanisms discussed above under Relationship Trust address the Access Security dimension of the Accessibility Data Quality category. Reputation systems under Authority Trust address the reputation dimension of the Intrinsic Data Quality category. Audits and regulatory compliance mechanisms under Regulation Trust address the believability and objectivity dimensions of the Intrinsic Data Quality category.

MARKET FRICTIONS

Market friction involves anything that prevents markets from developing and working properly, and that imposes costs or restraints on business transactions [TheFreeDictionary —Frictionless Markets"; Designer Carrots]

Several major kinds of market friction arise from the uneven supply of information in the market [McMillan p44]:

- Search Costs: time, effort and money spent by buyers trying to learn what is available from sellers. In electronic data markets today it is often difficult to know what is available
- Evaluation Costs: difficulties buyers have in assessing product quality. Markets malfunction when buyers cannot observe quality
- Access Costs: difficulties buyers have in obtaining access to online digital products. Semantic and technical differentials between buyers and sellers must be overcome
- Transaction Costs: difficulties buyers have in using online digital products. Entropy happens: If not monitored and maintained, data will degrade over time

Search Costs

It takes time, effort and resource for buyers to learn what data products are available where and at what price. Buyers are empowered by anything that makes it easier for them to acquire information about their targeted acquisition [McMillan p46]. However, search costs tend to lock buyers into sellers and can allow sellers to overcharge. A lowering of search cost can bring a disproportionate lowering of selling prices. Any market innovation that lowers search costs or makes search more effective will make markets more efficient: a) people waste less time and money on search, b) better matches of buyer and seller are formed, and c) pricing becomes more competitive to the buyers advantage. Search tools help the user to identify value-added information that is beneficial and provides advantages from their use, as well as data that is relevant (appropriate and useful to a particular task) [Wang].

Some of the traditional mechanisms related to managing & sharing product knowledge are useful here. There is a long list of market devices that aid the acquisition of information and mitigate the anticompetitive effects of the cost of search [McMillan p45]. Consumer Reports, Yellow Pages, word-of-mouth, advertising, loss-leaders, brand names, trademarks, market intermediaries all fall into this category.

Search costs in many cases have been dramatically lowered by the internet [McMillan p19]: Online Catalogs – descriptive information about products, services, and suppliers

- Search Tools – for finding product information along various dimensions
- Blogs – provide buyer experiences with products & sellers
- Shopping Robots (Bots) – Electronic agents that automatically search online merchants to find best price goods that meet specified requirements

- Specialized comparison shopping services for given products

Search being a wasteful activity, buyers might be willing to pay for an intermediary service that conducts the search for them [McMillan p45]. Information brokers provide this role in online data markets.

Evaluation Costs

Customers can have a difficult time assessing whether the product they are considering buying is worthwhile [McMillan p50]. Buyers need assurances of various kinds: that the product will be delivered on time, that the product will be in good, usable condition, that problems with the product will be addressed in a timely manner, and that there are no security risks associated with acquiring or using the product. Significant transaction costs derive from difficulties in observing quality. Channels of information that allow buyers to verify the quality of what they are acquiring can increase trust. Payment schemes can then be introduced under which prices reflect measured quality. Brand names, authoritativeness designations, or certifications can then be introduced to give buyers trust in what they are receiving.

A prime example of this is in markets in personal data. There is a need, given the current situation, for a market trust infrastructure that will assure users that their personal data will not be revealed and that collection is secured at the primary service provider. Further assurance involves the fact that secondary market aggregation and mining processing has security, integrity and validation built into it [Kiayias]. Even for black markets in data, lack of trust among the criminals themselves can depress prices. It can be very uncertain to do business with stolen data. Buyers struggle to determine if information is fresh, firsthand, authentic, or if it will even be delivered or paid for as promised. Sellers often worry that a payment from a hijacked PayPal or bank account could be traced by cops [Sutherland].

Various quality assurance mechanisms are available for exploitation to serve this area. Service Level Offerings (SLOs) have been proposed as a way to publicize or expose information about the level of quality of a particular data asset. The SLO can represent a current, average or guaranteed quality level. It can be structured to address multiple data quality dimensions as needed: accuracy, precision, completeness, consistency, timeliness, pedigree, etc. Inclusion/access to quality information can be provided along with other product metadata during the search and discovery processes. Automated profilers can be used to measure and report product quality at each stage of the distribution chain. These are especially valuable if they can be operated by suppliers during the creation and maintenance of information products and included in product registration as part of the service level offering (SLO). Automated profilers or quality checking engines can also be operated by the customers against the information products or samples of the product, possibly as cloud validation services. Finally, appraisal companies, operating as intermediaries, can offer buyers, for a fee, appraisals of the goods by experts [McMillan p79].

Access Costs

After having found a data asset, and determining it to be acceptable, it can also be difficult for customers to actually gain access to the data. There are a number of reasons for this. There may be semantic inconsistency where the semantics in the expression of what the customer needs may not match the terminology & semantics of the data sources. There may be format & syntax inconsistency where there are significant technology differentials between the buyers and sellers operational environments and the data assets to be exchanged.

Overcoming access problems can become a very involved and costly process. The use of tools or resident facilities is often adequate. The avoidance of point-to-point solutions is critical. The application of standards is very advisable. The involvement of acquisition resources for design & development is frequently necessary. The reuse of existing capability becomes extremely valuable. Frequently, suppliers, customers or intermediaries will need to construct a customized access mechanism for reconciling

differences between the suppliers' environments and data constructs, and the customers' environments and usage constructs in order to select and provide exactly the right data

A variety of traditional mechanisms are frequently available to access the data [Day]. Search and browse tools for drilling into, isolating and extracting specific content of interest. Charting tools used for displaying data trends after discovery through the browser. Data dumps permit the desired data to be downloaded to operate on locally.

On the other hand a public web API and/or web query language can be provided by a data market for developers to access their data and operate on the data in their servers [Day]. If a market doesn't wrap its data via an API, it's making things difficult for the developer. When used in the context of web development [Wikipedia ―Application programming interface"], an API is typically a defined set of Hypertext Transfer Protocol (HTTP) request messages, along with a definition of the structure of response messages, which is usually in an Extensible Markup Language (XML) or JavaScript Object Notation (JSON) format.

For example [Day], the Yahoo! query language (YQL) is an SQL-like query language created by Yahoo! as part of their Developer Network. YQL is designed to retrieve and manipulate data from APIs through a single Web interface, thus allowing mashups that enable developers to create their own applications. On the other hand JSON is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate. It is based on a subset of the JavaScript Programming Language, Standard ECMA-262 3rd Edition - December 1999. JSON is a text format that is completely language independent but uses conventions that are familiar to programmers of the C-family of languages, including C, C++, C#, Java, JavaScript, Perl, Python, and many others. These properties make JSON an ideal data-interchange language.

One practice is publishing web APIs to allow the combination of multiple services into new applications known as mashups. The practice of publishing APIs has allowed web communities to create an open architecture for sharing content and data between communities and applications. In this way, content that is created in one place can be dynamically posted and updated in multiple locations on the web.

While "Web API" is virtually a synonym for web service, the recent trend (so-called Web 2.0) has been moving away from Simple Object Access Protocol (SOAP) based services towards more direct Representational State Transfer (REST) style communications. Data markets providing their data through a RESTful API are referred to as providing "Data as a Service" (DaaS) or "cloud data".

Another type of market mechanism available to address access issues is international standards (e.g., ISO 22745 & 8000 standards related to master data quality). Adoption and implementation of these standards can provide for high quality levels in information assets involved in ongoing exchanges. They involve the use of formalized representations of catalogs that allow the capture and management of master data (used to construct individual transactions) in a structured and systematic way. They also provide for Open Technical Dictionaries that formalize and exploit the semantics and terminologies used in transactions, and Identification Guides that provide syntax schemes & business rules that formalize and exploit the syntactic and formatting requirements for exchanges

Finally, when a customized access solution absolutely must be constructed, and the development of any needed interfaces must be performed by the buyer, the seller or an intermediary, then it is critical to ensure the incentives for all the participants are properly established [Fichman, Leach]. Otherwise unneeded or unnecessarily complex interfaces and infrastructure might be constructed, easily reusable assets may not get created and exploited, and/or cost and schedule advantages of reusing existing approaches and solutions may be forgone. Creation of custom access solution will frequently initiate

participation in an entirely separate, acquisition market where contractors bid to provide needed access mechanisms

Transaction Costs

Successfully completing the data exchange transaction can itself be fraught with challenges and hidden costs. One thing that many data markets lack is the ability to buy a small subset of data from premium datasets [Day]. Most markets enable you to purchase premium datasets using a credit card, falling back to a phone call if the default option doesn't work. This is fine for large scale purchases such as entire multi-hundred-dollar datasets, although the credit card processing fees can be steep for both seller and data market provider. However, this breaks down when one wants to sell smaller datasets, or even per-use slices of data from a dataset. Credit card processing fees are simply not cost-effective for small payments.

Another transactional issue is how data markets fit into the overall management of information supply chains. The state-of-the-practice is simply inadequate [Seligman]. For example, in order to initially incorporate a purchased asset into an existing information supply chain, and then periodically, regularly or continuously receive updates to the information product, a frequently long and complicated long development and integration activity must be employed. Any proposal to change any of a number of physical and/or operational characteristics of the product must typically run the gamut of rigid configuration management boards with long approval and change cycles. In other cases there are only informal, ad hoc agreements between suppliers and consumers with no supporting technical infrastructure. And the problems will only get worse. Because organizations are increasingly building value-added services on top of data resources they do not control, any changes in semantics, representation or quality levels can wreak havoc to the business processes they support. It is critical to monitor these transactions to understand how well expectations on performance and behavior are being met.

Different transactional mechanisms are available or can be envisioned to address these particular problems. With regard to payments, there are several options to streamline the payment process and enable more flexible means of payment [Day]. A pay-per-dataset scheme can be introduced for the customer who doesn't want to buy the whole dataset, but only a chunk that they have drilled down to that is particularly relevant to their need. Micropayment mechanisms are available for making payments of very small amounts. Micropayment tools have allowed a data market to lower transaction costs to the point where they are now financially reasonable for smaller purchase sizes, thus increasing profit for the market and its data providers. Embedded payment mechanisms provide a visual presentation of the payment flow in which the sender appears to never leave the checkout or payment page. Embedded payments make it easier for a sender to make a payment by allowing the sender to bypass the payment login step. Various means of aggregating payments can be enabled which make the purchase of digital goods easier and can reduce fees through the use of micro pricing (special rates for low cost goods). All of these schemes open the door to service level pricing whereby payment schedules can be introduced under which prices reflect graded and measured quality levels.

For real-time, periodic or frequently accessed data sources, production monitoring mechanisms will be needed to regularly evaluate & provide feedback on the quality dimensions deemed important by the buyer. These will ensure that expected or agreed upon quality levels are being met, and do not change suddenly or degrade over time. These monitoring mechanisms would feed into another mechanism called data sharing agreements. Data sharing agreements (service level agreements for data) provide a new kind of service level agreement specifically for data that provides the wizards for agreement specification, and a technical infrastructure that supports monitoring and enforcement, change impact analysis, and data flow analysis (e.g., who are the critical providers) [Seligman]. The expected benefits of data sharing agreements are greater clarity through explicit obligations, increased data sharing through increased levels

of trust between consumers and providers, and reduced administrative burdens (e.g., "piggybacking").

Quality Dimensions Related to Overcoming Market Frictions

With respect to the hierarchical framework of data quality, several of the categories and dimensions come into play in overcoming market frictions through the application of various mechanisms.

First there are particular quality related aspects of search that should be considered. Search benefits users when it exhibits high recall and high precision. Search should address the value-added, relevancy and possibly most of the other dimensions of the Contextual DQ category. A high recall, high precision search returns all the information and only information relevant to a particular task (search request) [Wikipedia "Recall & precision"]. Completeness refers to the degree to which the values are present in the collection and the data are of sufficient breadth, depth and scope for the task at hand.

Many of the mechanisms used in the evaluation step can be configured and applied to address the accuracy dimension of the Intrinsic DQ category, and possibly the timeliness, completeness and amount of data dimensions of the Contextual DQ category.

Access should address all of the dimensions of the Representational DQ category, as well as the access dimension of the Accessibility DQ category. For transaction processing, the implementation of international standards as well as the application of various audits and controls can significantly improve quality levels while reducing overall transaction costs.

CONCLUSION

The Trusted Data Market must first establish a foundation of Trust. This foundation must include having a governing authority that is respected by all parties to do the right thing and operate the market in a reliable, consistent, fair and efficient manner. The trust foundation should also ensure diligence in monitoring operations to ensure compliance by all parties, and address grievances that are encountered by market participants, as well as promote high levels of confidence in the successful and satisfactory conclusion of transactions within the marketplace. Given a solid trust foundation, the trusted data market must then implement a series of market mechanisms to ensure smooth and efficient market operations by reducing friction (many related to data quality) in the areas of search, evaluation, access and transaction processing.

FUTURE DIRECTIONS

There are significant opportunities to provide buyers and sellers with mechanisms that facilitate the evaluation of the quality of information products. Some areas requiring significant work include: tools to do the evaluation work, standards for communicating evaluation results, devices (e.g. dashboards and display widgets & wizards) for reaching agreements and conducting transactions. These mechanisms are foundational to improvement in all of the friction areas, and advances in the maturity of the trust model. Furthermore, ways are clearly needed to connect these new evaluation capabilities to the other friction areas and to support a fuller realization of the trust model. Future work is needed to better understand the differences in application of Trusted Data Market principles to different types of data markets (real-time feeds, free public data sources, internal data markets, etc.) and the particular types of needs they may present.

Future work is also need to more fully relate the DQ body of knowledge organized through the framework for data quality (and the data quality management techniques that have evolved to address specific data quality dimensions), to trusted data markets. This paper only provides a basic mapping. Case studies that explore their application through the trusted data markets framework to address specific

analyzed trust problems and market frictions as well as the measurement and impact analysis of data quality management in this regard would be a fruitful path for investigation.

REFERENCES

[1] Bailey, J. P., *Intermediation and Electronic Markets: Aggregation and Pricing in Internet Commerce*, PhD Thesis, MIT, May 1998, http://*citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.90.1434&rep=rep1&type=pdf*.

[2] Birnbaum, B., Jaffe, A., *Relational Data Markets*, Principles of DBMS, Class Project – Fall 2007, CSE 544 Web: © 1993-2011, Department of Computer Science and Engineering, University of Washington, www.cs.washington.edu/education/courses/544/09wi/project/.../alex-ben.pdf.

[3] Burth, A., E-Management Markets: Transforming Organizations Beyond Stovepipes, 2006, ISBN, 978-1-4303-0490-6.

[4] BusinessDictionary, ―Price.‖ *BusinessDictionary.com – Online Business Dictionary,* WebFinance, Inc*.,* 2010, *as retrieved on January 13, 2011,* http://www.businessdictionary.com/definition/price.html.

[5] DACS, The Data Analysis Center for Software, Copyright © 2010, QSI – Quanterion.com, https://goldpractice.thedacs.com/practices/arrc/

[6] Day, B., ―Selling Digital Goods in Data Markets‖, PayPal Developer Network, Copyright © 1999-2011 PayPal, https://www.x.com/docs/DOC-3294/

[7] ―Designer Carrots, Market-Based Instruments for NRM Change‖, Australian Government, as retrieved on April 22, 2011, http://www.marketbasedinstruments.gov.au/Home/tabid/36/Default.aspx.

[8] DSE, Data Services Environment, US Department of Defense, Defense Information Systems Agency (DISA): https://metadata.ces.mil/dse/about.htm

[9] Fichman, R.G., Kemerer, C.F., *Incentive Compatibility and Systematic Software Reuse*, Journal of Systems and Software, Vol. 57, Issue 1, Apr. 2001, http://www2.bc.edu/~fichman/Fichman_2000_Incentives.pdf

[10] Forsell, M., Halttunen, V., Ahonen, J., ―Use and Identification of Components in Component-based Software Development Methods‖, Software Reuse: Advances in Software Reusability, 6th International Conference, Vienna, Austria, June 2000, Lecture Notes in Computer Science 1844, Frakes, W. (Ed.), Springer, Berlin.

[11] Fukuyama, F., *Trust: The Social Virtues & The Creation of Prosperity*, 1995, The Free Press, Simon & Schuster, Inc., New York, NY.

[12] Gislason, H., *The Emerging Field of Data Markets – Our Competitive Landscape*, DataMarket Blog, DataMarket © 2008–2011, http://blog.datamarket.com/2011/02/25/the-emerging-field-of-data-markets-our-competitive-landscape/.

[13] Hubbard, W.H., *How to Measure Anything: Finding the Value of "Intangibles" in Business*, 2007, John Wiley & Sons, Inc., Hoboken, NJ.

[14] Juran, J.M., Juran's Quality Handbook, Fifth Edition, 1999, McGraw-Hill, New York, NY, http://www.ee-techs.com/adm/quality-handbook.pdf.

[15] Karhoff, H., ,Buck, K., *Air Force Depot Maintenance System (AFDMS) MRP II/MRO Investment Analysis AFMC/LGN*, MITRE Technical Report, MTR 02B,Spring 2002, MITRE Washington C3 Center.

[16] Kiayias, A., Yener, B., and Yung, M., *Privacy-Preserving Information Markets for Computing Statistical Data*, Financial Cryptography and Data Security, Lecture Notes in Computer Science, Volume 5628. ISBN 978-3-642-03548-7. Springer Berlin Heidelberg, 2009, p. 32

[17] Klein, R., *Hurricanes and Residual Market Mechanisms,* May 29, 2009, Center for RMI Research, Georgia State University.

[18] Koronios, A., Redman, T., Gao, J., *Internal Data Markets: The Opportunity and First Steps*, Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology, 2009, IEEE, 978-0-7695-3898-3/09.
[19] Leach, R.J., *Software Reuse: Methods, Models and Costs*, McGraw-Hill, 1997, ISBN 0070369291
[20] Maryland Office of Peoples Council, *What Are Energy Suppliers? Brokers? Aggregators?,* retrieved January 11, 2011 from http://www.opc.state.md.us/LinkClick.aspx?fileticket=70JejiEyw84%3d&tabid=71
[21] McMillan, J., *Reinventing the Bazaar: A Natural History of Markets*, 2002, W. W. Norton & Company, New York, NY.
[22] Metron Aviation, Inc. , *Market Mechanisms for Airspace Flow Program Slots, NextGen Airspace,* 2009, SBA Proposal, Proposal Number: 09-2 A3.01-8820.
[23] Myers, F., *Acquisition M&S Master Plan, A Progress Report to DS System-of-Systems Architecture Modeling Review*, September 22, 2005, PowerPoint briefing by the Principal Assistant OUSD (AT&L) DS/SE/DT&E and Chair Acquisition M&S Working Group.
[24] Reifer, D.J., *Practical Software Reuse*, Wiley, 1997, ISBN 0471578533
[25] Seligman, L., Swarup, V., *Data Sharing Agreements (a.k.a., "SLAs for Data")*, MITRE Technology Program, internal MITRE Corporation briefing, April 19, 2005.
[26] Simba Information, *Business Information Markets 2005-2006*, Published: Sep 1, 2005, http://www.simbainformation.com/sitemap/product.asp?productid=1576984
[27] Surowiecki, J., The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations, 2004, Doubleday, Random House Inc., New York, NY.
[28] Sutherland, B., *The Rise of Black Market Data*, Newsweek, Dec 15, 2008, URL: http://www.newsweek.com/2008/12/05/the-rise-of-black-market-data.html .
[29] TheFreeDictionary, —Fictionless Market.", *TheFreeDictionary,* Farlex, Inc., Copyright © 2011, as retrieved on 22 April 2011.
[30] U.S. Congress, Office of Technology Assessment, *Remotely Sensed Data: Technology, Management and Markets,* OTA-ISS-604, Washington D.C.: US Government Printing Office, September 1994.
[31] Van de Loo, K., *Organizational Impact of Enterprise Services Architecture (ESA)*, 21 Dec 2004, Copyright © 2005 SAP AG, Inc., https://www.sdn.sap.com/irj/servlet/prt/portal/prtroot/com.sap.km.cm.docs/library/web-application-server/m-o/Organizational%20Impact%20of%20Enterprise%20Services%20Architecture%20ESA.article
[32] Wang, R.Y., Strong, D., *Beyond Accuracy: What Data Quality Means to Data Consumers,* Journal of Management Information Systems, 1996, 12(4), pages 5-34.
[33] Wikipedia, —Application programming interface." *Wikipedia, The Free Encyclopedia*. Wikimedia Foundation, Inc., as modified on 29 April 2011 at 23:20, and retrieved on 3 May 2011.
[34] Wikipedia, —Credit Bureau." *Wikipedia, The Free Encyclopedia*. Wikimedia Foundation, Inc., as modified on 21 December 2010 at 10:16, and retrieved on 14 January 2011.
[35] Wikipedia, —Information security." *Wikipedia, The Free Encyclopedia*. Wikimedia Foundation, Inc., as modified on 28 April 2011 at 17:59, and retrieved on 29 April 2011.
[36] Wikipedia, —Market." *Wikipedia, The Free Encyclopedia*. Wikimedia Foundation, Inc., as modified on 27 November 2010 at 01:56, and retrieved on 27 November 2010.
[37] Wikipedia, —Market clearing." *Wikipedia, The Free Encyclopedia*. Wikimedia Foundation, Inc., as modified on 1 October 2010 at 17:23, and retrieved on 13 January 2011.
[38] Wikipedia, —Market system." *Wikipedia, The Free Encyclopedia*. Wikimedia Foundation, Inc., as modified on 19 January 2011 at 00:54, and retrieved on 24 April 2011.
[39] Wikipedia, —Precision and recall." *Wikipedia, The Free Encyclopedia*, Wikimedia Foundation, Inc., as modified on 1 April 2011 at 15:24, and retrieved on 4 May 2011.