Information Aggregation - A Value-added E-Service

Paper 106

June 2001

Hongwei Zhu
Michael D. Siegel
Stuart E. Madnick

To appear in the *Proceedings of the International Conference on Technology, Policy, and Innovation*: *Critical Infrastructures*, The Netherlands, June 26-29, 2001.

Information Aggregation – A Value-added E-Service

Hongwei Zhu, Michael D. Siegel, Stuart E. Madnick
MIT Sloan School of Management
30 Wadsworth Street
Cambridge, MA 02139
{mrzhu, msiegel, smadnick}@MIT.EDU

Abstract

Information aggregation, a service that collects relevant information from multiple sources, has emerged to help individuals and businesses to effectively use the growing amount of information on the Web. In this paper, we analyzed a number of characteristics of information aggregation, namely comparison, relationship, and intra-organization aggregation. Using a case study, we demonstrated that successful e-business models could be built using information aggregation. The key to success is to identify value creation mechanisms using the aggregated information and combining it with domain knowledge of a specific industry and market. With these mechanisms, value-added post-aggregation services can be provided to generate sufficient revenues to sustain and grow the business. Emerging policies and regulations will impact the development of information aggregation. Although it is technologically feasible to aggregate information globally, the differences in aggregation related policies among nations may become a barrier and harmonizing them still remains a challenging task.

Keywords: aggregation, business model, e-business, e-service, policy

1. Introduction

The amount of information from conventional databases and web sources has been growing exponentially in the past few years. How to effectively use this massive amount of information has become a challenge to individuals and organizations. While existing businesses are making adaptive transformations to cope with the dynamics of the digital economy, new technologies and services are emerging to help integrate and extract values from relevant information sources.

This study focuses on a particular type of service that aggregates information from multiple sources using Internet technologies. After introducing the notion of information aggregation, we identify its major characteristics and illustrate each with examples of existing service providers. Enabling technologies for information aggregation will be discussed briefly.

Most of the existing studies on e-business models tend to cover a spectrum of business types, lacking in-depth analysis of the details that determine the viability and sustainability of e-businesses (Timmers, 1998; Mahadevan, 2000; Rappa, 2001). We will take a more focused approach and use a case study on a business in information aggregation to demonstrate that the key to success comes from the aggregated information and the capability of providing value-added e-services.

Information aggregation also raises a number of policy and regulatory issues. We will describe these issues and different approaches taken in the U.S. and Europe. With the trend of the integration of world economy, information aggregation will operate globally. The challenge of formulating effective international policies that nurture the growth of electronic commerce will be discussed.

2. Information Aggregation

2.1. Definition of Information Aggregation

Information aggregation is a service that gathers relevant information from multiple sources to provide convenience and add value by analyzing the aggregated information for specific objectives using Internet technologies. We call the providers of this service aggregators.

In a broader sense, information intermediaries such as newspapers, magazines, professional journals, and more recently, increasing number of web portals are information aggregators since they all collect information from multiple sources and disseminate it for convenient consumption. But they are not included in this study because that they tend to serve the general needs for information and lack the functionality of analyzing the aggregated information at different levels of granularity for specific goals.

Similar to the definition for information aggregation in an earlier study (Madnick, *et al.*, 2000), the definition in this paper focuses on a subset of information intermediaries to allow for in-depth analysis of business, technology, and policy related issues. This definition will become clearer when we discuss its unique characteristics and unfold its value creation mechanisms in the following sections.

2.2. Characteristics of Information Aggregation

Like generic information intermediaries, web aggregators collect, categorize, and regroup information from multiple sources. In addition, they perform analysis to the aggregated information. Based on a survey of over a hundred existing and emerging web aggregators, an earlier study (Madnick, *et al.*, 2000) categorized web aggregators according to their capabilities. Here we will refine these notions with further explanations.

2.2.1. Comparison Aggregation

Nowadays the enormous amount of information on the web has made it a difficult task to search for specific information. In the case of online purchasing, a search engine offers little help in finding a product and a competitive vendor that carries the specific product. For example, a search for “palm pilot” using the most popular search engine, Google, returns about 483,000 URLs! It will be a laborious task to visit each of these sites and find the pertinent information.

Comparison aggregation has emerged to address this problem. It retrieves information about a set of attributes for a product/service offered by many competing vendors and normalizes the information for meaningful side-by-side comparison. Many shoptbots and search agents, such as mySimon.com and DealTime.com, offer comparison aggregation services.

The attributes being compared are product/service specific. For example, for books, a comparison aggregator may retrieve the merchant name, merchant rating, title and format of the book (e.g., hardcover, audio tape, etc.), availability, price, total price after tax plus shipping and handling, etc. For services such as wireless communication, it may retrieve information such as provider, plan name, whether there is a promotion, minutes included, monthly fee, long distance rate, etc. To ensure that the information is meaningful, aggregators should reconcile semantic conflicts among different sources. For instances, DealTime.com can recognize currency inconsistencies and make correct conversions using the latest exchange rate.

2.2.2. Relationship Aggregation

To attract customers and promote customer loyalty, many companies are offering convenient online services, such as banking and brokerage. This has in turn greatly increased consumer choices among different service providers. Owning a close relationship with their customers is important for service-oriented businesses.

From a consumers' perspective, they bear the burden of managing their relationships with multiple service providers. It is common for a person to have multiple bank accounts, brokerage accounts, and credit cards. Since each account requires a login, it would be tedious and time consuming to manage those accounts.

Fortunately, the emergence of relationship aggregation can help customers manage multiple accounts with a single logon. A relationship aggregator can collect the information on a customer's behalf and generate various useful reports. The convenience of accessing all information from one place helps establish a close relationship between the aggregator and its customers. In the U.S., a number of financial account aggregators have been developed to help customers manage disparate accounts, e.g., Yodlee, VerticalOne, and Corillian, to name only a few. Realizing the potential threat of loosing customers and opportunities of cross selling, many financial institutions are now offering aggregation services to their customers. Web portals such as AOL and Yahoo are also offering account aggregation.

2.2.3. Intra-organization and Inter-organization Aggregation

It is common that the information within an organization is distributed among information systems run by different departments and branches. An intra-organization aggregation service can aggregate relevant information from disparate sources to promote knowledge sharing and perform firm level analysis. We will give an example of this type of aggregation in the case study.

Intra-organization aggregation can also be used as an alterative to system integration and standardization in some large organizations. For example, integration between old systems and new ones can be done through aggregation with minimal development cost. A similar situation exists of inter-organizational aggregation, as well as mixtures of intra- and inter-organization information being aggregated.

2.3. Enabling Technologies

Many aggregators do not have any arrangement with information sources and therefore can only extract information from web sites that are accessible publicly or by using customer's identity. As we know, HTML based web pages are semi-structured for display within a browser. Traditional database management tools such as SQL query processors cannot retrieve information from web pages. Web wrapping technologies such as Cameleon (Firat, *et al*., 2000) have been developed to sift through web pages and extract specified information elements. Using web wrapping technology, aggregators can retrieve information from web sources as if they were traditional databases. This sometimes has been called "screen scraping" in the press. Its performance can be slow when there are a large number of sources and it will also depend on the availability of underlying source web sites. Many aggregators use local databases to cache retrieved information in order to improve performance.

Extensible Markup Language (XML) has been developed as a better way to describe data for information interchange and content management. An XML-based industry standard, Open Financial Exchange (OFX), has been developed by CheckFree, Intuit, and Microsoft to facilitate accurate and secure financial information interchange. But it requires the participating institutions to provide information that conforms to the standard. Unfortunately it is often difficult to achieve this agreement. Currently, most financial account aggregation is still done through screen scraping.

In some cases, aggregators can pre-arrange with sources to receive a direct data feed from them. Any mutually-agreed encoding standard can be used, although XML is now most preferable. Further processing is still required to facilitate various aggregation services.

In addition to the capability of retrieving information from a large number of heterogeneous sources, aggregators should also be able to detect and reconcile semantic differences among autonomous information sources. This requires knowledge of metadata information of each source, including conventions and assumptions being used. One can implement conversion programs using the metadata knowledge. However, this tightly coupled approach does not scale nor can it gracefully handle semantic changes in sources. When dealing with complex conflicts among large number of sources, a loosely coupled approach developed in the Context Interchange (COIN) framework (Madnick, 1999) seems to be suitable to achieve a high level of scalability and robustness.

3. Case Study: Intra-organization Aggregation

Here we will use a case study to show how an aggregator can choose appropriate enabling technologies and combine a number of aggregation characteristics to offer value-added post-aggregation services.

Cadence Network (www.cadencenetwork.com) is an intra-organization aggregation service provider based in Cincinnati, Ohio. The company aggregates a variety of maintenance, repair, and operation (MRO) information and non-core business expense information for multi-location enterprises, and makes the aggregated information accessible through its password protected web site.

Many companies operate facilities that are geographically dispersed, especially those in the retail sector, e.g., lodging, grocery, video rental, and various chain stores. The number of facilities for such a company usually ranges from a few hundred to a few thousand across North America. Each facility independently chooses the vendors for services such as electricity, gas, water, sewer, solid waste, telecom, etc. While each individual bill seems to be small, they add up quickly and may eventually affect the profit margin of the company. Managing these non-core business expenses has been a traditional challenge given the large amount of scattered information and complex pricing structures from multiple vendors. This is the exact problem that Cadence Network helps to solve.

Cadence obtains information from its strategic business partners. By linking to Cass Information Systems and Insite Services, the two largest utility bill payment processors in the US, Cadence has direct access to detailed information on the actual consumption, pricing, billing, and payment of various services at each facility. Cadence also partners with Encompass to aggregate detailed MRO information. Its aggregated HVAC preventative maintenance data allows customers to monitor their systems in one place and any inefficient rooftop unit can be pinpointed for timely repair.

Figure 1 shows the information aggregation process of Cadence. Since the information from Cadence's partners has already been highly aggregated, Cadence can be viewed as an aggregator of aggregators, also called a mega-aggregator (Madnick, *et al.*, 2000). Information interchange between Cadence and its partners uses mutually agreed encoding standards. Upon receiving information from partners, Cadence normalizes it into its local data store for the purpose of performance improvement and integration with other internal information. By serving many customers, Cadence has the benefit of economies of scale.

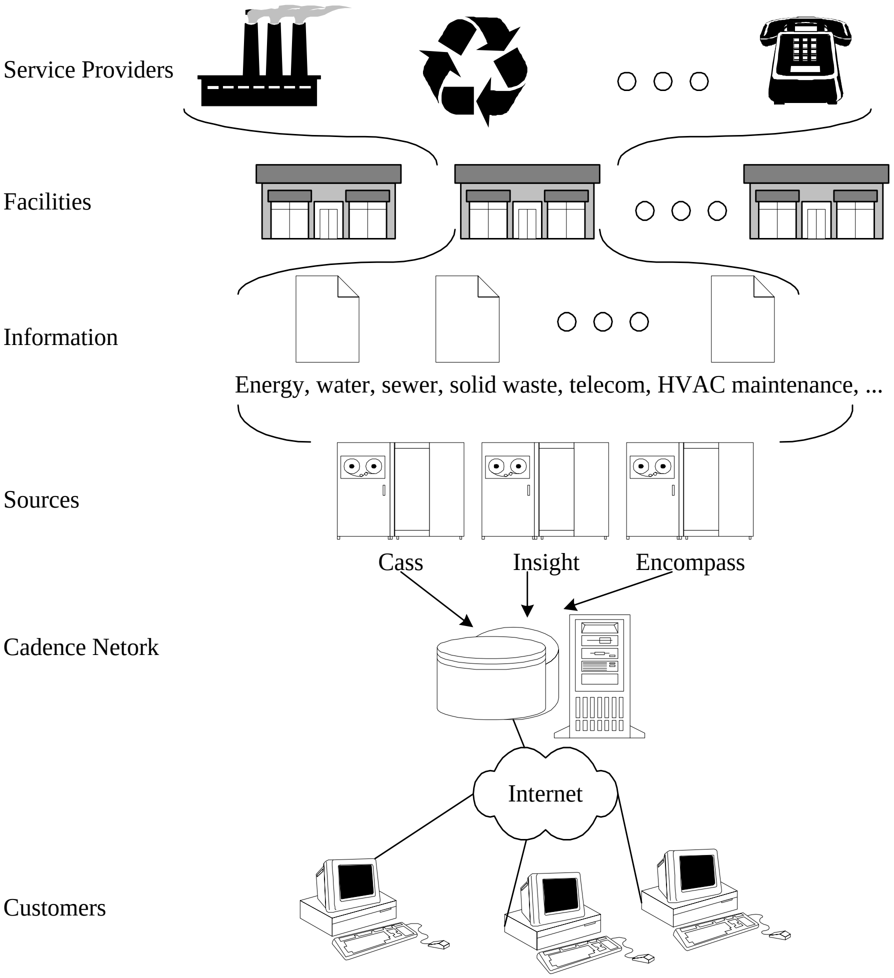
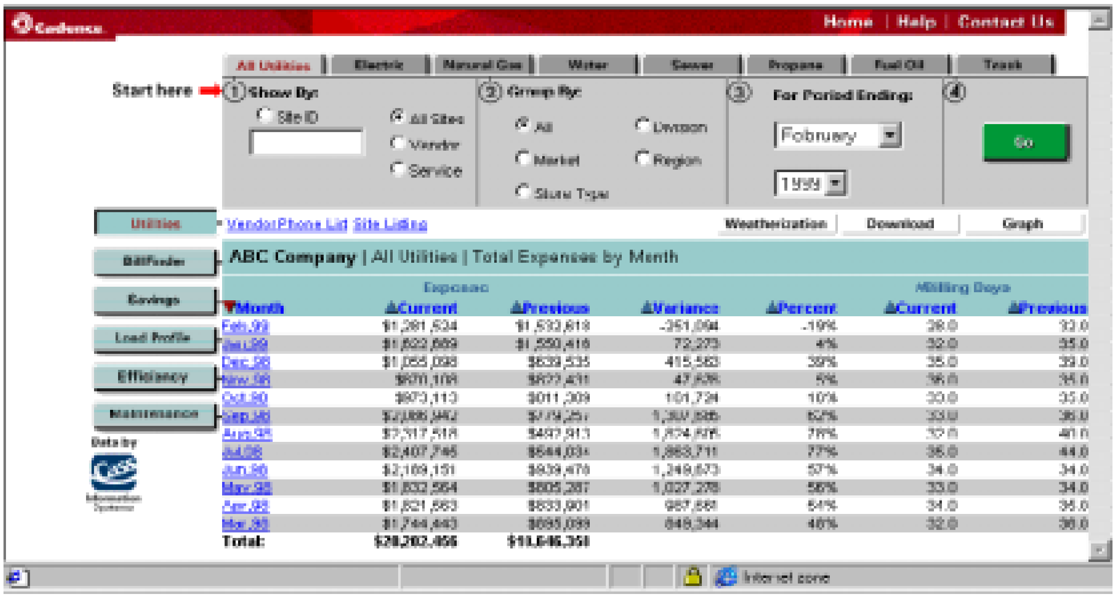


Figure 1. Information Aggregation Process at Cadence Network, Inc.



ABC Company | All Utilities | Total Expenses by Month

	Expense				Billing Days	
Month	Current	Previous	Variance	Percent	Current	Previous
Feb.99	$1,281,524	$1,532,618	-251,094	-19%	28.0	32.0
Jan.99	$1,622,689	$1,550,416	72,273	4%	32.0	35.0
Dec.98	$1,055,098	$639,535	415,563	39%	35.0	39.0
Nov.98	$870,109	$822,431	47,678	5%	36.0	35.0
Oct.98	$973,113	$871,389	101,724	10%	33.0	35.0
Sep.98	$2,086,942	$779,251	1,307,691	62%	33.0	36.0
Aug.98	$2,317,518	$492,913	1,824,605	78%	32.0	40.0
Jul.98	$2,407,745	$544,034	1,863,711	77%	35.0	44.0
Jun.98	$2,189,151	$939,478	1,249,673	57%	34.0	34.0
May.98	$1,832,564	$805,287	1,027,278	56%	33.0	34.0
Apr.98	$1,821,583	$833,901	987,682	54%	34.0	35.0
Mar.98	$1,744,443	$895,099	849,344	48%	32.0	38.0
Total:	$20,202,456	$10,646,358				

Figure 2. Sample Screen Shot of Aggregated Information

The raw information created and collected by Cadence partners is for the purposes of generating bills and processing payments. Cadence aggregates the information for completely different purposes, i.e., providing the convenience for their customers to view/manage all facilities in one place; and facilitating information analysis for cost reduction. The Web-based reporting tool offered by Cadence has rich functionality to fulfill these goals. Figure 2 gives an example report and the user interface that a Cadence customer can use within a Web browser. Managers from a multi-location enterprise can use the tool to analyze overall cost; cost by service type, division and other grouping mechanisms; total purchase from each service provider; details of each site; etc. The comparison tool is especially useful. It allows for comparisons among different facilities and comparisons between different time periods for the same facility. These comparisons are useful to identify inefficient facilities and predict future cost. They can also be used for benchmarking cost reduction measures. It is worth noting that for energy related comparisons weather variations can be normalized to get a comparable baseline.

A portion of Cadence's revenues comes from subscriptions to the information aggregation service. With knowledge about the customers and the market, Cadence also provides a variety of post-aggregation services. Each of these value-added services generates a revenue stream for the company.

Bill auditing. Billing errors are difficult to spot when information is scattered and unorganized. Once information is put into context using the reporting tool, those errors become easier to detect. Cadence will contact the service provider for refund arrangement on behalf of their customers and claim a fraction of the refund as commission. One may think that billing errors are rare with modern computerized billing systems. However, the result from Cadence is quite surprising. They have detected many billing errors, the largest single error being as high as $45,000.

Cost effective procurement. It is common that service providers have pricing schemes that favor volume purchasers, a form of secondary price discrimination. Traditionally, decentralized procurement practices in multi-location enterprises disqualify them for any volume discount. With Cadence's service, these enterprises now can use their aggregate demand information to negotiate with vendors to get a better rate. Oftentimes customers lack the knowledge of the market. Cadence's experienced business experts can assist or represent customers in negotiation. The impact is substantial. In one case Cadence saved its client $1.5 million with the knowledge of customer's aggregate demand and their proficient understanding of complex tariffs. Cadence receives a commission for this type of service.

Other advisory services. Cadence is staffed with knowledgeable business analysts who are also experts in a specific industry, e.g., energy, telecom, etc. They constantly analyze new development and best practices in their specialized industry. Their knowledge is also augmented by studying information of their customers from all over the country. Knowing both customer needs and what the market can offer puts them in a credible position for providing fee based advisory services. For example, their rate analysis can help diagnose a customer's rate relative to national or regional average. They can also help identify energy inefficiencies and recommend improvement plans.

All of these services are based on one or more mechanisms that can create values to Cadence and its customers. Table 1 summarizes these mechanisms and values.

Table 1. Value Creation Mechanisms in Cadence's Information Aggregation Services

Mechanism	**Value**
Information aggregation	High efficiency due to economies of scale; convenience of seeing all information in one place
Reporting and analysis tools	Managing multiple remote facilities; informed budgeting and planning; problem diagnosis
Aggregate demand	Cost reduction
Comparisons	Identifying errors and problems; measuring effectiveness; projecting with trend; cost reduction with rate analysis
Knowledge of both customer and market	Credible and valuable advices

4. Business Models for Information Aggregation

Using cases of comparison aggregation and relationship aggregation, Madnick, *et al.* (2000) identified a number of business models for consumer oriented information aggregation services. They envisioned that the aggregated information has tremendous value and aggregators should look into ways of extracting the value to provide rich post-aggregation services. The case study on Cadence confirms that post-aggregation services can be the primary value adding mechanism for business oriented information aggregation.

The key to value creation in post-aggregation service is the knowledge about the customers as well as the market. In the Cadence case, they know their customers very well by looking into the aggregated information. This helps them determine the customer needs and identify problems. Cadence also has the advantage of accessing information of all customers. This helps them to study the overall market and enrich their domain knowledge. For example, looking at only one customer's information does not help to diagnose if the customer is paying a rate higher than industry or national average for the same service. With the knowledge of both customers and the market, Cadence can perform such evaluations and effectively reduce cost by giving advices or helping customers to negotiate a competitive rate.

Like other Internet companies (Fischer, 2000), aggregators can choose between selling aggregation enabling product and providing services as their business model. In the Cadence case, a service model is adopted because of the complexity of information aggregation. The upfront setup cost of the system is intimidating to many companies in retail industry. These companies also lack the technological sophistication required for maintaining the system. In addition, business analysts at Cadence are domain experts who can effectively use the aggregated information to deliver highly value-added services. It is sensible for Cadence to choose the service provider model. By contrast, if an aggregator does not have the capability of providing a variety of post-aggregation services (e.g., due to lack of domain knowledge, customer trust, etc.) and subscription alone does not generate sufficient revenue, it may need to seek other business models, such as licensing its aggregation technology or hosting aggregation for other service providers. Recently some financial account aggregators licensed their technology to financial institutions for them to provide aggregation services. These aggregators essentially become technology or application service providers. And from customers' perspective, financial institutions become aggregators. This phenomenon also serves as the evidence of the importance of post-aggregation services to a business in information aggregation.

Depending on the characteristics of information aggregation, an aggregator may choose different business models. The key to sustainability of an information aggregation business is value creation mechanisms that generate sufficient revenue streams. Cadence's model clearly demonstrates this point.

Although it is out of the scope of this paper, aggregators should also adopt a pricing strategy for their services in order to maintain their competitiveness while extracting maximum consumer surplus.

5. Policy Issues for Information Aggregation

Information aggregation is raising a number of issues that are not specifically addressed in the old legal and policy regime. For example, some financial aggregators can make transactions between accounts but they have not been regulated as financial institutions in the U.S., therefore the liability of any transactional errors will not be clear until the Electronic Fund Transfer Act (i.e., Regulation E) is amended to address this issue. Relationship aggregation involves significant amount of personal data, which will result in serious privacy concerns that need to be addressed.

Two other controversial issues, property rights in cyber space and intellectual property rights over data, have been raised in a number of lawsuits filed in the past couple of years against aggregators (e.g., eBay vs. Bidder's Edge, mySimon vs. Priceman, First Union vs. Paytrust, and Ticketmaster Corp. vs. Tickets.com Inc.) Some of the aggregators have stopped parts of their screen scraping practices to avoid litigation, and some did go to the court. In the case of eBay vs. Bidder's Edge, the latter being an aggregator that systematically queries online auction information from the web sites of eBay and other online auction shops, the court issued a preliminary injunction based on "trespass to chattels" concept found in old common law. Whereas in the Ticketmaster vs. Tickets.com case, the court rejected the trespass claim because entering a publicly accessible site should not be considered as trespassing.

As to the issue of intellectual property for databases, data collected in databases is not protected under the current copyright law since factual information is not attributed to "original works of authorship". But for economic reasons, databases should receive certain protection so that creators have enough incentives to create and maintain databases. Conversely over-protection will restrict fair use and free flow of information, which also threats the development of information aggregation. To address this issue, two controversial bills were introduced in the Congress in 1999: The Collections of Information Antipiracy Act (HR 354), and the Consumer and Investor Access to Information Act of 1999 (HR 1858). HR 354 has much stricter restrictions. Although neither bill was passed, they indicated a trend that databases will be protected. If over-protection is issued, aggregation that uses "screen scraping" technologies without approval of the sources may face some legal challenges.

Different approaches have been adopted in different countries to deal with emerging issues. The U.S. government recognized the importance of the Internet to the economy and has been promoting its development by adopting five flexible policy principles (The White House, 1997), the core concept of which is that the government should "avoid undue restrictions on electronic commerce", and when government involvement is needed, it should help to create a "predictable, minimalist, consistent, and simple" legal system. As a result, there have been few Internet related regulations in the U.S. Electronic commerce in the U.S. has benefited from the flexibilities granted by this policy framework and thrived in the past few years.

The European Union (E.U.) has taken a completely different approach and issued a number of strict regulations to govern the development of electronic commerce. The E.U,'s Data Protection Directive, enacted in 1998, grants the creators of databases exclusive right of protecting the data from unauthorized extractions. Databases of a non-E.U. company will not be protected under this directive unless its home country adopts the same or a very similar law. The E.U. has been using this reciprocity-based provision to induce other nations to follow the lead of this regulation. This has created some pressure to the U.S. and resulted in the introduction of the two aforementioned database protection bills (Lee and McKnight, 1999).

The E.U. has also introduced a directive to safeguard the privacy of personal information. They have announced the intent of blocking data flows into and out of countries that do not provide adequate protection. The U.S. privacy law has been diverse (i.e., many types of privacy laws, each having a specific context not designed with the Internet in mind) and decentralized (i.e., both federal and state laws) (Glancy, 2000). With this status quo of privacy law, U.S. has been reluctant to take a centralized approach and more inclined towards an alternative of self-regulation by industry (Samuelson, 1999).

As the world economy becomes more integrated, more businesses will be conducted globally, which requires free flow of information among different systems across country borders. This is quickly becoming technologically possible with the fast growth of the Internet and services like information aggregation. But it may take quite some time to overcome the barrier resulted from the differences in policies of different countries and regions. Since each country may seek to preserve the uniqueness of its social values during the globalization process, it is very difficult to harmonize legal standards using reciprocity-based rules. In this case, it will probably work well if nations can achieve "policy interoperability" by agreeing on the goals that a policy should achieve, and letting each nation decide the implementation details of the policy (Samuelson, 1999).

6. Conclusion

We have seen dramatic growth in the amount of information on the Web. This trend will continue in the future as the last-mile bottleneck is being removed in developed countries and infrastructure is put in place in developing nations. Finding relevant information and extracting value from it is becoming more important for businesses and individuals. The emergence of information aggregation on the Internet provides an effective way of retrieving and managing relevant information that is dispersed all over the Web. The opportunities abound for businesses to provide value added services using aggregated information.

Information aggregation is useful in a number of ways, such as comparing goods and services, providing personalized service in exchange for close relationship, and gathering information from different parts of an organization. These characteristics can often be combined to maximize values in the aggregated information. For instance, a relationship aggregator that aggregates financial accounts can also aggregate information on various investment instruments to assist customers to adjust their investment portfolios. In the case of Cadence Network, the intra-organizational information aggregation service provider also offers effective comparison analysis tools to its customers.

Value creation mechanisms are important for the success of e-businesses in information aggregation. By extracting value from the aggregated information and combining it with domain knowledge in a specific industry, an information aggregator can provide a variety of value-added services, which will generate multiple streams of revenue necessary for the sustainability and growth of the business.

Policy issues are also arising and will have some impact on information aggregation. Data protection should balance between protecting the investment in data creation and promoting value added data reuse and free flow of information. As information aggregation operates more globally, harmonizing policy differences among nations will become a major challenge for service providers as well as policy makers.

Acknowledgement
The study has been supported, in part, by BSCH, Fleet Bank, Merrill Lynch, and Suruga Bank. The authors would like to thank Mr. Mark Shunk, who was CEO of Cadence Network when the case study was conducted, for his help and support.

References

1. Firat, A., Madnick, S., Siegel, M. (2000) "The Cameleon Web Wrapper Engine", Proceedings of the VLDB Workshop on Technologies for E-Services, Cairo, Egypt.
2. Fischer, L.M. (2000) "Product or Service? Internet Infrastructure's Battling Business Models", *Strategy+Business*, Issue 21, 79-87.
3. Glancy, D. (2000) "At the Intersection of Visible and Invisible Worlds: United States Privacy Law and the Internet", 16 Santa Clara Computer and High Technology Law Journal 357.

4. Lee, T, McKnight, L. (1999) “Internet Data Management: Policy Barriers to an Intermediated Electronic Market in Data”, 27th Annual Telecommunications Policy Research Conference.
5. Madnick, S.E. (1999) “Metadata Jones and the Tower of Babel: The Challenge of Large-Scale Semantic Heterogeneity”, 1999 IEEE Meta-Data Conference, April 6-7, 1999.
6. Madnick, S.; Siegel, M.; Frontini, M.A.; Khemka, S.; Chan, S.; Pan, H. (2000) “Surviving and Thriving in the New World of Web Aggregators”, MIT Sloan Working Paper #4138.
7. Mahadevan, B. (2000) “Business Models for Internet based E-Commerce: An Anatomy”, *California Management Review*, 42(4).
8. Rappa, M. (2001) “Business Models on the Web”, http://ecommerce.ncsu.edu/business_models.html.
9. Samuelson, P. (1999) “Five Challenges for Regulating the Global Information Society”, University of California at Berkeley.
10. Timmers, P. (1998) “Business Models for Electronic Markets”, *EM-Electronic Market*, 8(2), 3-8.
11. The White House (1997) A Framework for Global Electronic Commerce, http://www.ecommerce.gov/framewrk.htm.